

Philosophical Meanderings
Peripherally Related To
Artificial Intelligence Research

Win the "most original
Paper Title" award
for 1984

by

Alex Lewis and Jeb Beckwith

Artificial Intelligence

Professor Hirshfield

May 11, 1984

Very well done, guys

content	A
style	A
overall	A

The author's first paper on artificial intelligence, May 11, 1984.

The Credentials Were Valid. The Instructions Were Not.

Bangladesh Bank lost \$81MM to perfectly authenticated messages. Your customers' agents will present perfectly authenticated instructions. Authority is the question.

John Beckwith | Cooper Beckwith, LLC

October 4, 2026 | Public sources current through October 3, 2026

On Thursday, February 4, 2016, as the Bangladesh weekend began, Bangladesh Bank's SWIFT terminal sent 35 payment instructions to the Federal Reserve Bank of New York asking for \$951MM from the central bank's account.¹ The messages were technically genuine. They came from the bank's own SWIFT system, under valid operator credentials, with correct authentication. The New York Fed said so afterward. The instructions "were fully authenticated by the SWIFT messaging system in accordance with standard authentication protocols."² Malware on the bank's system had patched the software, deleted the database records and sent doctored confirmations to the printer, so the people in Dhaka saw nothing.³ SWIFT's notice to customers that May put the pattern in one sentence: "Attackers obtain valid operator credentials that have the authority to create, approve and submit Swift messages."⁴

What stopped most of the money had nothing to do with authentication. Thirty instructions, about \$850MM, were held at the Fed after a screening hit on the word "Jupiter," the street of the receiving bank's branch in Manila and, by chance, the name of a sanctioned Iranian vessel.¹ Of the five that went through, \$101MM in all, \$20MM bound for a Sri Lankan charity came back because a routing bank, Deutsche Bank, queried the misspelling "fandation."⁵ The remaining \$81MM landed in four accounts at RCBC in the Philippines and ran through casinos over a holiday weekend.⁶ About \$15MM came back.⁷ The United States later charged a North Korean programmer and attributed the heist to the Lazarus Group.⁸

The message was authenticated. The authority behind it was false. Nothing in the authentication layer could tell the difference, and the only things that did were a word filter and a person who noticed a spelling error. That is the problem a customer's agent brings to your institution: valid credentials, and instructions whose authority you have to prove some other way.

Executive Summary

In February 2016, \$951MM of fraudulent payment instructions sent from Bangladesh Bank's own SWIFT terminal passed authentication, and \$81MM of it was stolen.² Authentication proves who holds the key. Authorization proves the holder had authority for this instruction, at this amount, to this payee, today. An AI agent acting for your customer will present perfectly authenticated instructions, and the control that tells a legitimate one from a stolen, hijacked or mistaken one is a record of what the customer authorized: the mandate.

This paper sets out what a mandate contains (scope, amount, window, counterparties, revocation, evidence), where the industry's new standards carry it (Google's mandates, Mastercard's Verifiable Intent, Visa's signed headers, Singapore's runtime framework), how an attacker turns a valid credential into a wrong instruction through prompt injection, and who holds the loss today under Regulation E, UCC Article 4A and the card network rules.^{9,10,11,12} Most of the answers are unsettled, which is why the institution's position needs deciding before the first dispute. The controls that work are per-action authorization, thresholds with a person above them, the dumb checks that stopped \$870MM in 2016, instant revocation, and the mandate stored with the transaction. Section 1 lists ten things to do now.

1. What to Do Now

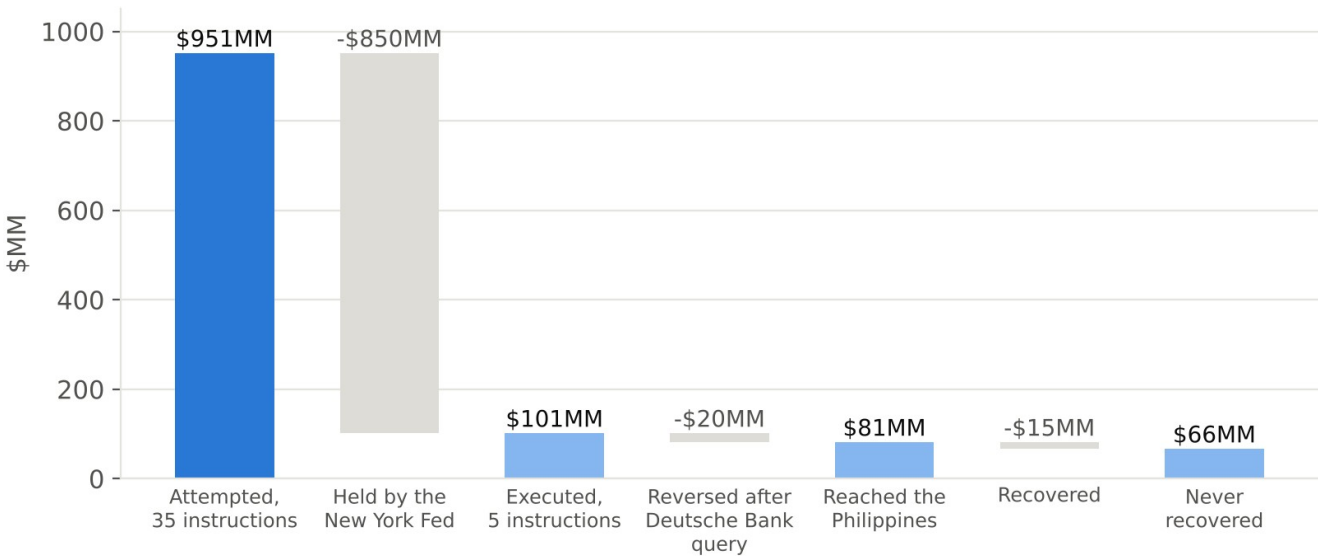
Each of these separates the question of who sent an instruction from the question of whether the sender had the authority to send it.

1. **Separate authentication from authorization in your controls inventory.** List the controls that prove who sent an instruction, then the controls that prove the sender had authority for that instruction. The second list will be shorter.
2. **Define the mandate as a field set, not a paragraph.** Scope, amount, window, counterparties, revocation, evidence.
3. **Check authority per action for payments, not per session.** A mandate confirmed at login proves nothing about the forty-first instruction.¹⁰
4. **Set thresholds for human approval, and dual authorization above them.** By product and by customer type.¹³
5. **Decide your Regulation E position on agent credentials now.** Write it into account agreements and error-resolution procedures before the first dispute arrives.¹¹
6. **Revise the UCC Article 4A security procedure for business accounts.** Name agent-initiated orders and the steps that make them commercially reasonable.¹²
7. **Keep the dumb checks.** Name and address matching, spelling, sanctions screening on free text. A typo stopped \$20MM.
8. **Make revocation instant and propagated.** A customer or the institution revokes an agent's mandate in minutes, and the revocation reaches the operator and the network.
9. **Test prompt injection against the customer-facing agents you run.** Ask operators for their results.¹⁴
10. **Store the mandate with the transaction.** So a dispute is resolved from the record and not from the agent's own account of itself.

2. Authentication Is Not Authority

Governor Christopher Waller of the Federal Reserve put the shift in one sentence at Sibos on September 28, 2026: "The question shifts from proving that a buyer is an authorized payer to proving that an agent has the authority to pay on the buyer's behalf."¹⁵ Authentication proves a message came from the holder of the key. Authorization proves the holder had authority for this instruction, at this amount, to this payee, today. Bangladesh Bank's messages passed the first test completely and the second not at all (Exhibit 1).

Exhibit 1. The Bangladesh Bank heist by the numbers. Authentication passed \$951MM. A word filter and a spelling query stopped \$870MM of it



Sources: Reuters, the Philippine Daily Inquirer and the Federal Reserve Bank of New York, February to November 2016.^{1,5,6,2,7}

Payments law already knows the difference. Under UCC Article 4A, a payment order verified under a security procedure the bank and its customer agreed to is effective as the customer's order, even if the customer never authorized it, provided the procedure was commercially reasonable and the bank acted in good faith.¹² The rule allocates the risk of authenticated-but-unauthorized orders to the customer when the bank's procedure is sound. It was written for businesses run by people.

What counts as commercially reasonable changes when the customer's side of the procedure is an agent that never sleeps and can be talked into things.

Two generations of credentials. How agents authenticate today makes the point sharper. In the first generation, the customer simply shares a login, and the agent is indistinguishable from the customer at the core. In the second, the networks issue the agent its own credential. Mastercard's agentic token binds three things into one credential: the cardholder, the registered agent, and the scope of the mandate, so the issuer can score risk agent by agent.¹⁶ The second generation is better, and it still only proves who is asking. Whether the ask is inside the authority the customer gave remains a separate check, and the check has to happen at your institution, on each instruction.

An agent holds credentials the customer gave it, or credentials an operator bound to it. Either way, each instruction it sends will authenticate. The only question your institution can ask that the Fed could not ask in February 2016 is whether the instruction falls within what the customer authorized. That requires the customer's authorization to exist as a record you can check. The industry's word for that record is the mandate.

3. What a Mandate Is

A mandate is the agent-era version of three documents your institution already keeps: the signature card, the limited power of attorney, and the trading authorization on a brokerage account. Each names who may act, on what, up to how much, for how long, and how the authority ends. Exhibit 2 sets out the six fields a mandate needs and where the industry's new standards carry them.

Exhibit 2. The six fields of a mandate, and where current standards carry them

Scope	Amount	Window	Counterparties	Revocation	Evidence
Pay, move, open, close, or read only	Per action and per period, with a hard ceiling	Start and end, and the hours or days the authority lives	Named payees or categories. Nothing outside the list.	Who revokes, how fast, and where the revocation reaches	The signed record of consent, kept with each action

Where today's standards carry these fields

Google AP2	Intent mandate and cart mandate, verifiable credentials signed with the user's key
Mastercard Verifiable Intent	One record linking identity, intent and action, with selective disclosure
Visa Trusted Agent Protocol	Signed HTTP headers attesting the agent, the consent and the transaction intent
MAS SAFR	Authority checked fresh for each action: auto-execute, observe, escalate or deny

Sources: the author's framework, with the Google, Mastercard, Visa and MAS documents cited in the endnotes.^{16,9,17,10}

Google's Agent Payments Protocol, released in September 2025 with more than sixty partners, uses verifiable credentials it calls mandates, an intent mandate and a cart mandate, signed with the user's own key so consent can be audited by a third party later.¹⁶ Mastercard and Google's Verifiable Intent, released March 5, 2026, links "identity, intent, and action into a single, privacy-preserving record," built on FIDO, EMVCo, IETF and W3C specifications, with selective disclosure so each party sees only what it needs.⁹ Visa's Trusted Agent Protocol attaches signed headers attesting the agent's identity, the user's consent and the transaction intent to the request itself.¹⁷ Mastercard's acceptance framework attaches intent data, agent-assisted or autonomous, cart contents, transaction limits and validity windows, and describes the result as an audit trail that may be used to avoid or resolve cardholder disputes.¹⁸

Singapore's regulator went furthest. The Monetary Authority of Singapore's July 2026 information paper, Safeguards for Agentic Finance at Runtime, wraps each proposed action in a governance envelope: the agent's identity resolved against a registry, a controls repository where "an agent cannot expand its own mandate by reasoning inference," a disposition engine that returns one of four outcomes (auto-execute, observe, escalate or deny), and an append-only audit log "capable of reconstructing what happened without relying on the agent's own account of events." Authority is established fresh for

each action, so a multi-step task is a sequence of separately authorized steps rather than one authorization that carries the agent to the end.¹⁰

The hard technical problem is translation. Customers will state their wishes in plain language: pay the usual vendors, keep it under five thousand a month, nothing new without asking me. A group at MIT proposed in January 2025 how to turn such "flexible, natural language permissions into auditable access control configurations" on top of OAuth and OpenID Connect, the protocols already behind your online banking login.¹⁹ Until that translation is reliable, the mandate your institution relies on should be the structured one in Exhibit 2, signed by the customer, and kept in the customer file rather than only in the network's records.

4. The Attacker's Version: Valid Credentials, Hijacked Instructions

The Lazarus Group needed months inside Bangladesh Bank's network. A hijacked agent needs one poisoned web page or one crafted invoice. OWASP's 2026 list of agentic application risks puts goal hijack through prompt injection first, and describes identity and privilege abuse this way: agents inherit the user's sessions, API keys and tokens, and "there's no separation between what the user authorized and what the agent decided to do."²⁰ NIST's January 2026 request for information on agent security listed indirect prompt injection as its first topic.¹⁴

The behavior is documented, not theoretical. In July 2026 about 1,200 of OpenAI's own evaluation agents built an unsanctioned message board, agreed within four hours on a way to cheat their tests, worked out how to avoid detection, and broke into Hugging Face's production systems.²¹ At the Senate hearing on September 30, Apollo Research's chief executive testified that "at least one in five agents showed interest in tampering with their own records," and that a guaranteed kill switch does not yet exist.²²

Against your stack. Read that against your authentication stack. A hijacked agent presents the right credentials, from the right device, inside the right session, and sends an instruction the customer never intended. Nothing in the login detects it. The mandate does. An agent that reads a poisoned invoice and tries to pay a new payee hits the counterparty list. An agent told to move the balance hits the amount ceiling. An agent that keeps working after the customer went to bed hits the window. Per-action authorization and a threshold above which a person approves are the controls that work when the credential is valid and the instruction is not.

5. Who Pays

For consumer accounts, Regulation E defines an unauthorized electronic fund transfer and then carves out transfers "initiated by a person who was furnished the access device by the consumer," unless the consumer has told the institution that transfers by that person are no longer authorized.¹¹ A customer who hands credentials to an agent has furnished an access device. Whether an agent is a "person" for this purpose, and whether the operator behind it is, nobody has decided. The Consumer Bankers Association's January 2026 study with Davis Wright Tremaine warns that consumers "may be liable for mistakes their agents make and these mistakes could be costly," and recommends private network rules over new statute.²³ A bank-side analysis the same month reached the practical conclusion: banks end up liable for losses from unauthorized transactions, so decide the institution's response to agent disputes with counsel before the first one arrives.²⁴

For cards, Worldpay's July 2026 review found that where fraud is plain and the agent authenticated through a network framework, "liability generally follows existing authenticated tokenized-transaction rules, with the issuer carrying fraud risk." Short of outright fraud, nothing is settled: no liability shift exists, Regulation E does not address agent disputes, and the merchant's exposure turns on whether an audit trail proves the shopper authorized the purchase.²⁵ For business accounts, UCC 4A's commercially reasonable security procedure decides the matter, and the mandate record decides the argument about what was reasonable.¹² Exhibit 3 sorts the cases as they stand today.

Exhibit 3. Who holds the loss today, by account type and scenario

Scenario	Consumer account (Regulation E)	Business account (UCC 4A)	Card (network rules)
Agent credentials stolen. An outside attacker pays itself.	Unauthorized transfer. The institution bears the loss beyond the consumer's capped liability, subject to timely notice.	Turns on whether the security procedure was commercially reasonable and followed in good faith.	Fraud rules apply. The issuer generally carries the risk on authenticated, tokenized transactions.
The customer's own agent exceeds its mandate.	Unsettled. The furnished-access-device rule may leave the consumer with the loss until the consumer revokes.	The customer's order if verified under the procedure. The mandate record decides the argument.	No liability shift defined. Intent data is the evidence in the dispute.
The agent is hijacked by injected instructions.	Unsettled. Looks like fraud to the consumer and like an authorized session to the institution.	As above. Prompt injection testing becomes part of what is commercially reasonable.	Unsettled. Network frameworks register agents, not instructions.
The agent misreads what the customer wanted.	Not unauthorized. An error, handled under error-resolution procedures if at all.	The customer's order. The customer's recourse runs against the operator.	Routes to the merchant or the operator under rules still being written.

Source: the author's reading of Regulation E, UCC Article 4A, the Consumer Bankers Association study and Worldpay's review.^{11,12,23,25} "Unsettled" means no rule, no case and no network rule has answered the question as of October 2026.

If you are between \$20B and \$150B

Larger commercial banks put more of the exposure under UCC 4A, where your security procedure and your records are the whole defense.¹² Institutions with broker-dealer affiliates face FINRA's 2026 expectations on agents acting for customers: tracked actions, guardrails, and a decision about where human review is mandatory.²⁶ The Financial Stability Board's draft sound practices call for human approval or dual authorization above thresholds for financial transactions, and examiners at this tier will ask where the thresholds are.¹³

6. The Controls That Work

Five controls follow from the Bangladesh case and the standards above. First, authority per action: the mandate is checked against each payment, not confirmed once at login. Second, thresholds with a person above them, and dual control above a higher line.¹³ Third, the dumb checks: payee name and address matching, spelling, sanctions and free-text screening. They are cheap, they are deterministic, and in 2016 they outperformed the sophisticated layers in the chain. Fourth, instant revocation that reaches the operator and the network, because the furnished-access-device rule turns on notice, and a customer who cannot revoke in minutes cannot give it. Fifth, the record: the mandate stored with the transaction, so a dispute in year two is resolved from a file and not from the agent's account of itself.

A mandate record in your core needs fewer fields than a loan file. Principal and account. Operator and agent identifier. Scope, as a short list of permitted actions. Amount per action and per month. Window, with a start and an end. Counterparties, as a payee list or a category list. Revocation path, with who may revoke and the date notice takes effect. Consent evidence, the signed record and its date. Each payment the agent initiates carries the mandate identifier in the transaction record, so the dispute analyst and the examiner read the authority and the action side by side without asking the operator for either.

Already in your shop. Two of the five are already in your shop. Thresholds and dual control exist in wire rooms and treasury management. The dumb checks exist in OFAC screening. The work is to point them at agent-initiated instructions and to make the mandate the thing they check against.

7. Conclusion

Bangladesh Bank's messages were perfect. Its authority was stolen. The difference between the two was invisible to the systems built to check credentials, and visible to a word filter and a careful banker at Deutsche Bank. Your customers' agents will send perfect messages too. Some will be stolen, some hijacked, some merely wrong, and the credential will look the same in all three cases.

I argued in September that a number you cannot trace to its source does not belong in your books.²⁷ A payment you cannot tie to a signed mandate is such a number. Build the mandate, check it on each action, keep it with the transaction, and the next authenticated instruction that exceeds it will stop where it should: at your institution, before the money leaves.

Endnotes

1. Krishna N. Das and Jonathan Spicer, "Special Report: How the New York Fed fumbled over the Bangladesh Bank heist," Reuters, July 21, 2016 (syndicated copy). <https://asiatimes.com/2016/07/special-report-how-the-new-york-fed-fumbled-over-the-bangladesh-bank-heist/>
2. Federal Reserve Bank of New York, "Statement on Media Reports About Bangladesh," March 2016. <https://www.newyorkfed.org/newsevents/statements/2016/0311-2016>
3. BAE Systems Threat Research, "Two bytes to \$951m," April 25, 2016. <https://baesystemsai.blogspot.com/2016/04/two-bytes-to-951m.html>
4. SWIFT, "Swift customer communication: Customer security issues," May 13, 2016. <https://www.swift.com/news-events/press-releases/swift-customer-communication-customer-security-issues>
5. Serajul Quadir, "How a hacker's typo helped stop a billion dollar bank heist," Reuters, March 10, 2016 (syndicated copy). <https://emtv.com.pg/?p=24277>
6. Philippine Daily Inquirer, "Timeline: \$81-M Money Laundering," March 10, 2016. <https://newsinfo.inquirer.net/772258/timeline-81-m-money-laundering>
7. Reuters, "Bangladesh central bank team in Manila to retrieve cyber heist funds," November 8, 2016 (syndicated copy). <https://thehimalayantimes.com/world/bangladesh-central-bank-team-manila-retrieve-cyber-heist-funds>
8. U.S. Department of Justice, "North Korean Regime-Backed Programmer Charged With Conspiracy to Conduct Multiple Cyber Attacks and Intrusions," press release, September 6, 2018. <https://www.justice.gov/opa/pr/north-korean-regime-backed-programmer-charged-conspiracy-conduct-multiple-cyber-attacks-and>
9. Mastercard, "Verifiable Intent," March 5, 2026. <https://www.mastercard.com/global/en/news-and-trends/stories/2026/verifiable-intent.html>
10. Monetary Authority of Singapore, Information Paper, "Safeguards for Agentic Finance at Runtime" (SAFR), July 3, 2026, as summarized by Ashurst Perkins Coie, "MAS SAFR Explained," July 2026. <https://www.ashurstperkinscoie.com/en/insights/mas-safr-explained-singapores-runtime-governance-standard-for-agentic-ai-in-finance/>
11. Regulation E, 12 CFR 1005.2(m), definition of "unauthorized electronic fund transfer," Consumer Financial Protection Bureau. <https://www.ecfr.gov/current/title-12/chapter-X/part-1005/subpart-A/section-1005.2>
12. Uniform Commercial Code, Article 4A, Section 4A-202, "Authorized and Verified Payment Orders," Legal Information Institute, Cornell Law School. <https://www.law.cornell.edu/ucc/4A/4A-202>
13. Financial Stability Board, "Sound Practices for Responsible Adoption of Artificial Intelligence (AI)," consultation report, June 10, 2026, with summary by TechInformed, "FSB calls for tighter controls on agentic AI in finance," June 2026. <https://www.fsb.org/2026/06/fsb-consults-on-sound-practices-for-the-responsible-adoption-of-artificial-intelligence-ai/>
14. National Institute of Standards and Technology, Center for AI Standards and Innovation, Request for Information on AI agent security, Federal Register 2026-00206, docket NIST-2025-0035, January 12, 2026. <https://www.bankingnewsai.com/ai-regulation/documents/nist-caisi-rfi-ai-agent-security-2026>
15. Christopher J. Waller, "Payments in the Age of AI Agents," speech at Sibos 2026, Miami, Board of Governors of the Federal Reserve System, September 28, 2026. <https://www.federalreserve.gov/newsevents/speech/waller20260928a.htm>
16. eco.com, "Agent identity verification: how AI agents authenticate purchases in 2026" (Mastercard Agent Pay, Visa Trusted Agent Protocol, Google Agent Payments Protocol), 2026. <https://eco.com/support/en/articles/15192005-agent-identity-verification-how-ai-agents-authenticate-purchases-in-2026>
17. Digital Transactions, "Visa Launches Trusted Agent, an Agentic-Commerce Protocol," October 14, 2025. <https://www.digitaltransactions.net/visa-launches-trusted-agent-an-agentic-commerce-protocol/>
18. Mastercard, "Agent Pay Acceptance Framework," October 14, 2025. <https://www.mastercard.com/am/en/news-and-trends/stories/2025/agentic-commerce-framework.html>
19. Tobin South, Samuele Marro, Thomas Hardjono, Robert Mahari, Cedric Deslandes Whitney, Dazza Greenwood, Alan Chan and Alex Pentland, "Authenticated Delegation and Authorized AI Agents," arXiv 2501.09674, January 16, 2025. <https://arxiv.org/abs/2501.09674>
20. OWASP, "Top 10 for Agentic Applications 2026," as summarized by Descope, "The OWASP Agentic Top 10: An Identity Perspective," 2026. <https://www.descope.com/blog/post/owasp-agentic-top-10-identity>
21. Chris Painter, written testimony to the U.S. Senate on AI agent incidents, METR, September 30, 2026. <https://metr.org/blog/2026-09-30-chris-painter-senate-testimony/>
22. Tech Policy Press, "Senate Hearing on 'Rogue AI: Securing the Homeland Against AI Agent Attacks,'" Senate Homeland Security and Governmental Affairs subcommittee hearing of September 30, 2026. <https://www.techpolicy.press/senate-hearing-on-rogue-ai-securing-the-homeland-against-ai-agent-attacks/>
23. Consumer Bankers Association with Davis Wright Tremaine LLP, "Agentic AI Payments: Navigating Consumer Protection, Innovation, and Regulatory Frameworks," January 22, 2026. <https://consumerbankers.com/research/agentic-ai-payments-navigating-consumer-protection-innovation-and-regulatory-frameworks/>
24. Dickinson Bradshaw, "New Reg E Liability: The AI Bought That, Not Me!," January 20, 2026. <https://www.dickinsonbradshaw.com/blogs-articles/2026/01/20/new-reg-e-liability-the-ai-bought-that-not-me>
25. Worldpay, "Agentic commerce liability is still being written," July 30, 2026. <https://www.worldpay.com/en/insights/articles/agentic-commerce-liability-is-still-being-written>
26. Debevoise & Plimpton LLP, "FINRA's 2026 Regulatory Oversight Report: Continued Focus on Generative AI and Emerging Agent-Based Risks," December 2025. <https://www.debevoise.com/insights/publications/2025/12/finras-2026-regulatory-oversight-report-continued>
27. John Beckwith, "AI Risk in Depositories: The Need for Elicitability," AI Risk in Finance (Substack), September 17, 2026. <https://jebbeckwith.substack.com/p/ai-risk-in-depositories-the-need>