INSIGHTS · AI GOVERNANCE & CONTROLS

No Agent Should Audit Its Own Work

AI agents are moving into the financial close, and sometime in the next few audit cycles an external auditor is going to ask a question most companies cannot answer: how is the agent's output controlled? The answers on offer — "a person reviews it," "the model checks itself" — would not survive a walkthrough if a human performed the work the same way. Finance solved this exact problem decades ago and named the solution segregation of duties. The control transfers. The profession just hasn't noticed yet.

EvoNova Advisors | Insight | July 2026 | 9 min read

BY THE NUMBERS

95.5→89.0%

GPT-4's accuracy on a math benchmark before and after it was asked to review and revise its own answers. Self-review made the model worse.

Huang et al., ICLR 2024

~60%

how often two AI models that are both wrong agree on the *same* wrong answer. Shared blind spots are measurable — and highest between models from the same provider.

Kim et al., ICML 2025

73.5%

how reliably GPT-4 recognizes its own writing when asked to judge it — with self-recognition directly driving self-preference. The reviewer knows whose work it is grading.

Panickssery et al., NeurIPS 2024

The auditor's question is already scheduled

The agents are already in the building. PwC and Anthropic announced a collaboration this February aimed squarely at deploying enterprise AI agents in regulated industries, with finance named first: agents inside systems of record, doing liquidity forecasting and scenario modeling, wrapped in what the announcement calls "clearly defined human-in-the-loop controls." The major close-management platforms now ship agents that prepare reconciliations and match transactions inside SOX-relevant processes. The demand side tells the same story. Gartner puts AI usage at 59 percent of finance functions. Deloitte finds 54 percent of large-company CFOs naming agent integration a transformation priority, while a companion Deloitte study finds only about 14 percent have actually integrated agents into finance, with trust named as the main barrier. The gap between ambition and deployment is not a technology gap. It is a controls gap, and everyone involved can feel it.

The auditors can feel it too. In a 2024 PCAOB staff spotlight, audit firms told the regulator's staff that generative AI "could amplify certain existing information technology risks (e.g., risks related to the segregation of duties)." That is the audit profession connecting AI to SoD in a single sentence — in a regulator's publication. Protiviti's SOX practice reported last fall that external auditors have been extremely cautious about AI-performed work, and that companies should describe an approach that includes human review of all AI-generated work product, or expect substantive testing. The UK's audit regulator published guidance on generative and agentic AI in audits last June. The IIA refreshed its AI auditing framework. ISACA launched an AI audit certification. When the standard-setters, the external auditors, and the internal-audit profession all move within eighteen months, the question is not whether your agent controls will be examined. It is what you will say when they are.

Right now, the honest answer at most companies is some version of "a person skims it." I want to convince you that this answer fails, not because the person is careless but because it is the wrong control — and that the right control is one your own audit methodology has required of humans for decades.

You already know the answer. You wrote it down in 2000.

Strip the technology out of the question and look at what remains: work is being produced at volume by a fast, confident, occasionally wrong performer, and someone has to decide how that work gets checked before it matters. Finance has a name for this problem, and an answer older than SOX itself. Segregation of duties — the deliberate separation of who does the work from who checks the work — is a core control activity in the COSO framework, and the reason COSO gives is worth reading slowly: it reduces the risk *of error and fraud*. Not just fraud. Error. The maker-checker discipline that every bank back office runs was never only about catching thieves. It was about catching mistakes, by making sure the eyes that verify are not the eyes that produced.

The profession went further and codified *why* the same eyes cannot verify. The SEC's auditor-independence rule — Rule 2-01, on the books since 2000 — directs that independence is impaired by anything that would "place the accountant in the position of auditing their own work." The international ethics code calls it the self-review threat and lists it among the five fundamental threats to independence. Note what these rules do not say. They do not say the self-reviewer is dishonest. They say something more interesting: that a competent, honest professional evaluating their own prior judgment will inherit the assumptions and blind spots that produced it, and no amount of diligence undoes that. The control answer is structural, not motivational. You don't train the threat away. You separate the duties.

Everything about that logic transfers to AI agents. Almost nothing about current practice reflects it.

Self-review by machine has now been measured. It fails the same way.

What makes this moment unusual is that the machine-learning literature spent the last three years running, in effect, a controlled audit of self-review. It reached the same conclusion the accounting profession reached about humans, with numbers attached.

Start with the headline result. Researchers at Google DeepMind and the University of Illinois asked frontier models to review and revise their own reasoning with no outside information — the machine equivalent of "check your work before you submit it." Accuracy went down. GPT-4's score on a standard math benchmark fell from 95.5 percent to 89.0 percent over two rounds of self-correction; on a common-sense benchmark, an earlier model collapsed from 75.8 to 38.1 percent. The paper's sharpest finding is about the earlier studies that claimed self-correction worked: those results, it showed, had quietly used the answer key to decide when to stop correcting. Give the model an oracle and self-review looks great. Take the oracle away, which is exactly the condition of an agent preparing a reconciliation, and self-review subtracts value.

Follow-up work located the failure precisely. Models can *fix* errors: when researchers told models where an error was, they repaired it reliably. What models cannot reliably do is *find* errors in their own reasoning. The best model tested located the first mistake in its own chain of work barely half the time. Detection is the broken step, and that should sound familiar, because detection is the step segregation of duties exists to protect: the preparer who made the error is the person least equipped to see it.

And there is a mechanism underneath, one that maps directly onto the self-review threat. A NeurIPS study found that GPT-4 can recognize its own writing 73.5 percent of the time out of the box, and that the strength of this self-recognition predicts, almost linearly, how much the model favors its own output when judging. The recognition is causing the favoritism. A separate line of research showed that models sharing a conversation are subject to social pressure inside it: challenged on a correct answer, one production model wrongly capitulated 98 percent of the time. A reviewer that lives in the author's conversation inherits the author's context, anchors on the author's framing, and folds under the author's confidence.

> *The self-reviewer is not dishonest. It inherits the assumptions and blind spots that produced the work — and no amount of diligence undoes that. You don't train the threat away. You separate the duties.*

None of this requires the model to be badly built, any more than the self-review threat requires the accountant to be corrupt. It requires only that the checker share the maker's blind spots, and a model reviewing its own output shares all of them.

"Human in the loop" is answering the wrong question

The standard response is the phrase that shows up in nearly every AI governance deck and in the agent vendors' own announcements: human in the loop. A person will review what the agent produces. That sounds like a control. Under load, it is mostly a hope.

The problem is not the human; it is what sustained review does to humans, and the literature here is older than AI and brutally consistent. In a 2023 study in *Radiology*, experienced radiologists — fifteen-plus years of practice — read

mammograms alongside AI-suggested categorizations. When the suggestion was wrong, their accuracy fell from 82 percent to 45 percent. The machine's confidence became the reviewer's conclusion. This is automation bias, documented since the 1990s across aviation and clinical settings: given a mostly-reliable automated aid, people miss what the aid misses and approve what the aid asserts, and they perform *worse* than people working without the aid. Add the arithmetic of agents (output produced faster than any reviewer can meaningfully read, vigilance decaying measurably within the first half hour of monitoring duty) and "a person reviews everything the agent does" describes a queue, not a control.

Practitioners have started using an uncomfortable word for this kind of oversight: performative. The clearest recent illustration came from one global consultancy, whose internal review process waved through a government report containing AI-fabricated citations. The fabrications were caught by an outside academic with no stake in the deliverable, and the firm partially refunded the contract. The loop had humans in it. What it lacked was independence.

The original automation-bias research buried a finding that deserves more attention than the headline. When experimenters made reviewers *accountable* for outcomes — measured them, named them, made verification their job rather than their burden — automation bias dropped significantly. The active ingredient in effective oversight was never the presence of a human. It was an independent, accountable check. Finance has known this forever: we never controlled financial statements by having someone "keep an eye on" the preparer. We built a second, separate pair of eyes into the process and made its sign-off mean something.

So the question "should a human review agent output?" is badly posed. The controls question is the one SOX taught us to ask: *is the review independent of the preparation, and is someone accountable for it?* A human answer can fail that test; the preparer's own team skimming its own agent's output fails it daily. And a machine answer can pass it.

Independence is a dial, not a switch

If the reviewer needs to be independent rather than human, the design question becomes: independent *how much*? This is where an honest treatment has to concede something, because the obvious objection is correct as far as it goes: a fresh instance of the same model is not a different reviewer in the way a second accountant is. It shares the weights. It shares the training data. It will share some blind spots no matter how cleanly you separate the context.

The evidence says the objection is real — and priced. When two AI models are both wrong, they agree on the same wrong answer roughly 60 percent of the time, far more than chance, and the correlation is strongest between models from the same provider. A 2025 ICML study found that as frontier models get more capable, their mistakes are becoming *more* correlated, not less. That weakens every scheme where one model checks a similar one. But the same literature shows what independence buys at each level. Verification performed in a context that cannot see the original draft outperforms verification performed alongside it: in one controlled design, hiding the author's work from the checker more than doubled precision on a factual task. Panels of judges drawn from different model families track human judgment better than a single large judge, at a fraction of the cost. And before anyone concludes that shared blind spots discredit the whole idea, note that *humans* fail the pure-independence test too: the classic software-engineering experiment on the question found that programs written by 27 independent programmers failed together

far more often than independence would predict. We did not respond to that by abolishing the second reviewer. We layered controls.

So the design answer is a hierarchy, the same way we already tier review rigor to assertion risk:

EXHIBIT 1

The independence ladder — tier the reviewer to the risk of the assertion

Reviewer configuration	What the added independence removes	What remains — and where it fits
Same session self-check	Nothing. The checker inherits the author's context, framing, and errors — and defends them.	Not a control. Measured self-review made a frontier model *worse*.
Fresh context, same model	Contextual anchoring, in-conversation sycophancy, the author's framing and reasoning.	Weight-level blind spots remain. The floor: cheap, and it catches errors the authoring session cannot see.
Different model family	Provider-level error correlation — the ~60% same-wrong-answer overlap between similar models.	Still opinion-based review, and frontier-model errors are converging. Stronger, not sufficient alone.
External grounding	Opinion altogether: the checker recomputes the math, ties numbers to source systems, re-derives claims.	The strongest machine rung — limited to what can be recomputed or traced.
Sampled human review	Delegated accountability: a named person on defined escalation triggers, not a skim of everything.	Where accountability becomes personal and legal. The §302 signature never moved.

Source: EvoNova Advisors. Correlated-error and self-review findings per Kim et al. (ICML 2025) and Huang et al. (ICLR 2024); fresh-context effect per Dhuliawala et al. (ACL 2024).

A fresh context — same model, new session, no shared conversation history, no access to the author's reasoning — is the floor. It is cheap, it removes contextual anchoring and in-conversation sycophancy — though not the model's ability to recognize its own prose, which is one more reason this rung is a floor — and it catches a class of error the authoring session cannot see at all. A different model checks more; a different model *family* checks more still, because it breaks provider-level error correlation. External grounding — a checker that recomputes rather than rereads, ties numbers to source systems, reruns the math — is stronger than any opinion-based review. And sampled human review sits on top, where the accountability finally becomes personal and legal. That last rung is not optional decoration: SOX Section 302 still requires the CEO and CFO to personally certify the filing, and no architecture changes that. Segregation of duties never removed the accountable signer. It structured the review beneath the signature, which is exactly where agent-checks-agent belongs.

What this control looks like when it runs

This is not theoretical for me. My firm produces client deliverables with AI in the pipeline daily, and the segregation control runs on every one of them: the reviewing agent is never the authoring agent. The reviewer gets a fresh context, the work product, and the standards the work must be checked against — an evidence registry listing every quantitative claim with its source and a confidence grade. It explicitly does not get the authoring conversation, or the author's drafts and reasoning. It re-derives rather than re-reads: numbers tied back to sources, math re-performed, every claim traced to the registry or flagged.

What the fresh-context reviewer catches, again and again, are the errors the author could not see. A figure that quietly changed between page four and page seventeen, invisible to the session that wrote both pages because it "knew what it meant." A confident claim resting on a source that says something adjacent but not that. A statistic imported from an old draft whose vintage no longer qualifies it for a headline. Arithmetic the author repeated wrong twice, consistently; consistency is exactly what makes self-review useless against it. When the checking agent shares the author's context, it defends these errors. When it doesn't, it finds them. Same model. Different duties. That is one firm's log, not a published defect study — the field owes itself the study. But every catch is documented, which is more than a skim can say.

Two design rules do most of the work. First, the reviewer's output is a documented verdict, not a vibe: pass, fail, or escalate, claim by claim, with reasons. A control that leaves no evidence is not a control an auditor can test. Second, escalation to a human fires on defined conditions — disagreement between author and reviewer, claims above a risk threshold, anything touching a filing — so scarce human attention lands where the automation-bias research says it works: on a sample, with accountability, against specific claims, rather than skimming everything and absorbing nothing.

EXHIBIT 2

Anatomy of an agent maker-checker control

Prepare

The authoring agent produces the work plus an evidence trail: every quantitative claim tied to a source and graded.

Review independently

A fresh-context agent — no shared session, no author reasoning — re-performs the math and traces every claim.

Document the verdict

Pass, fail, or escalate — claim by claim, with reasons, logged. A control an auditor can actually test.

Escalate to a human

Defined triggers: author-reviewer disagreement, risk thresholds, filings. Sampled, accountable, specific.

Source: EvoNova Advisors, from the maker-checker control run in the firm's own AI delivery pipeline.

For a controller or an internal auditor, the encouraging part is how little of this is new. Write it down and it reads like any other control description: *the preparing agent's output is reviewed by an independent agent with no shared session state, against defined criteria, with results logged and exceptions escalated to named human owners.* That sentence would not startle anyone who has documented a SOX control. That is the point.

The objections, taken seriously

"SoD exists because people have incentives to cheat. Agents don't." Half right. The fraud rationale transfers weakly to a system with no motives — though prompt injection and compromised agents are busily restoring the adversarial case. But COSO's stated rationale was always error *and* fraud, and the error half transfers with full force. If anything it binds tighter: an agent's errors are systematic, so its self-reviews fail systematically too.

"Models are getting better at self-correction." They are; the honest title of the headline paper ends in "Yet," and newer reasoning models are explicitly trained to revise mid-stream. But controls are not built on capability forecasts. An uncontrolled self-check was not acceptable from a brilliant staff accountant, and her talent was never the issue. Independence was. An auditor cannot rely on a control whose operating effectiveness depends on the preparer having a good day.

"Regulators want humans overseeing AI, not more AI." Today, largely true. The audit firms interviewed for that PCAOB staff spotlight say human review of AI output remains essential, and the EU AI Act's Article 14 requires that high-risk systems be overseeable by natural persons. Nothing here argues otherwise. The independence hierarchy keeps humans exactly where regulation and common sense put them: accountable, at the top, certifying. What it removes is the fantasy that a human skimming everything is a functioning control layer. Machine-scale output needs a machine-scale first check, independent by construction, so that human review can be what the evidence says actually works: sampled, specific, and accountable.

Ask your next vendor one question

Agent adoption in finance is stalling on trust, and trust is not going to arrive as a feature. It arrives the way it always has in this profession: as controls someone can test. The companies that get agents into production inside regulated workflows will be the ones that can hand an auditor a control description, an independence rationale, and a log — not the ones with the most impressive demo.

The near-term moves are small. Inventory where agents already touch your close, and mark every point where the checking context is the making context; that is your deficiency list. Write the control description now, before the walkthrough, in the language you already use: preparer, reviewer, independence, evidence, escalation. Tier the independence of the review to the risk of the assertion, the way you already tier everything else. And when a vendor shows you an agent that "validates its own output," ask the question this whole article compresses into one line, and watch what happens: *who reviews the agent's work — and do they share a context?*

Your audit methodology answered that question for humans decades ago. It was right. Apply it to the machines.

An EvoNova Advisors Insight. Machine-learning findings are cited from peer-reviewed venues where available; preprints and vendor-affiliated research are labeled. Automation-bias and multiversion-programming studies are vintage-flagged and cited deliberately as foundational results. Forecasts are labeled as forecasts.

SOURCES

Huang et al. (Google DeepMind / UIUC) — *Large Language Models Cannot Self-Correct Reasoning Yet* — ICLR 2024

Kim, Garg, Peng, Garg (Cornell) — *Correlated Errors in Large Language Models* — ICML 2025

Panickssery, Bowman, Feng — *LLM Evaluators Recognize and Favor Their Own Generations* — NeurIPS 2024

Tyen et al. (Cambridge / Google) — *LLMs cannot find reasoning errors, but can correct them given the error location* — Findings of ACL 2024

Sharma et al. (Anthropic) — *Towards Understanding Sycophancy in Language Models* — ICLR 2024

Dhuliawala et al. (Meta AI) — *Chain-of-Verification Reduces Hallucination in Large Language Models* — Findings of ACL 2024

Goel et al. — *Great Models Think Alike and this Undermines AI Oversight* — ICML 2025

Verga et al. — *Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models* — 2024 *(vendor-affiliated research; directional)*

Knight & Leveson — *An Experimental Evaluation of the Assumption of Independence in Multiversion Programming* — IEEE TSE, 1986 *(vintage; classic human-independence result)*

Dratsch et al. — *Automation Bias in Mammography* — Radiology, 2023

Skitka, Mosier & Burdick — *Does automation bias decision-making?* / *Accountability and automation bias* — Int'l Journal of Human-Computer Studies, 1999–2000 *(vintage; foundational automation-bias studies)*

SEC — Rule 2-01 of Regulation S-X (Release 33-7919, 2000); Munter, *The Critical Importance of the General Standard of Auditor Independence* — 2022

IFAC/IESBA — *Exploring the IESBA Code, Installment 5: Independence* — self-review threat definition

PCAOB staff — *Spotlight: Generative AI in Audits and Financial Reporting* — July 2024 *(staff outreach summary; relays audit-firm observations)*

Protiviti — *Calmer Audits, Higher Bar: 2025 SOX Compliance Trends* — October 2025

FRC — *AI in audit: illustrative examples and documentation guidance* — June 2025

The IIA — *Artificial Intelligence Auditing Framework* (updated) — September 2024

ISACA — *Advanced in AI Audit (AAIA) certification launch* — 2025

PwC / Anthropic — *Enterprise agent deployment collaboration announcements* — February & May 2026

Gartner — *AI in Finance Survey* (183 CFOs) — November 2025

Deloitte — *Q4 2025 CFO Signals* (200 CFOs); *Trust as barrier to agentic AI in finance* — 2025–2026 *(two separate studies)*

Fortune / The Register — coverage of AI-fabricated citations in a consulting deliverable and partial contract refund — October 2025

EU — *Artificial Intelligence Act, Article 14 (Human Oversight)* — in force

An EvoNova Advisors Insight, synthesizing peer-reviewed machine-learning research with published guidance and surveys from the SEC, PCAOB staff, IESBA, FRC, the IIA, ISACA, Protiviti, Gartner, and Deloitte. Machine-learning findings are cited from peer-reviewed venues where available; preprints and vendor-affiliated research are labeled; vintage studies are flagged.