

HORIZON CREDIT ADVISORY · AGENTIC ASSURANCE

When the Credit Officer Is an Agent

Seamus Maher · Horizon Credit Advisory

Autonomous AI is arriving on trading-floor credit desks faster than anyone can govern it. The hard part is not what the machines get wrong. It is what they get right for the wrong reason, and who is left to notice.

A credit officer at a European gas and power trading desk opens her firm's daily counterparty review one October morning. A cold snap overnight has pushed regional gas prices up sharply, and one mid-sized producer's exposure has moved outside its normal range. Her firm's AI agent has already drafted a recommendation: no action, exposure within limit. The reasoning cites the producer's investment-grade rating.

The real reason the position is safe sits a few fields deeper in the same file. The trade is fully collateralised under a strong credit support annex, and the two protections answer different questions: a rating is a slow judgement about whether a name is likely to pay across the cycle, while collateral determines what is actually at risk on the morning the price moves. This morning the rating and the collateral happen to agree. They will not always, and nothing in the agent's own output would have told her which one was doing the work.

That gap, a correct recommendation resting on the wrong stated reason, is the governance problem now arriving on credit desks across energy and commodity trading. A wave of vendors sell autonomous "credit intelligence" agents that watch limits, exposure, margin events and counterparty news around the clock. They hand the officer a ranked set of recommended actions with the reasoning already written up. The promise is real: hours of officer time compressed into a defensible draft. The question underneath it has changed, though. It is no longer "can we automate the write-up?" It is "can we rely on, and defend, what the agent decided?"

THE FAILURE THAT OUTPUT CHECKS MISS

A right answer, resting on the wrong reason



Figure 1 · The same recommendation follows from either protection. Only one of them is doing the work on any given morning.

The model that will not sit still

Everything a risk function knows about validating models rests on one quiet assumption: the same inputs produce the same outputs. A potential future exposure engine, the simulation that estimates what a counterparty could owe you, is deterministic given its seed. Run it twice, get the same answer twice. That is precisely what lets you validate it, by reproducing the result and challenging the assumptions.

An agent built on a large language model breaks that assumption three ways at once. The same counterparty file can produce a different reasoning path, sometimes a different recommendation, from one run to the next. A trivial rephrasing of the dossier can flip the outcome. There is no fixed formula to document, so there is nothing to reproduce and reconcile.

Then there is drift, and drift is silent by nature. The vendor updates the model underneath the product, and the agent's behaviour shifts with no change ticket appearing anywhere in the firm's own systems. From the credit desk, nothing appears to have changed at all. And when specialist agents start interacting, one watching limits, another drafting recommendations, a third checking compliance, the failure modes that matter appear in how they combine, not in any single agent tested alone.

None of this has slowed adoption. Deloitte's 2026 State of AI in the Enterprise survey, covering 3,235 IT and business leaders across 24 countries, found only 21 per cent reporting a mature governance model for agentic AI. Roughly 80 per cent lacked the basic controls model risk management would take for granted: defined decision boundaries, real-time monitoring, a complete audit trail. By 2027, the same survey found, 74 per cent expect to be running AI agents at least moderately.

Nor is this a banking story that energy and commodity firms can watch from a distance. It is arriving inside the tools they already run. CubeLogic, whose RiskCubed platform is a credit-risk system built for energy and commodity trading firms, is releasing Version 12 of its Credit Risk product with embedded AI that lets users generate reports from natural-language prompts. A prototype in development reads unstructured counterparty financial statements and converts them into structured credit records, the exact task the opening scene turns on.

"CubeLogic is no longer the bottleneck for real-time credit risk; the constraint has moved to surrounding client systems," says Richard Winter, the firm's chief technology officer. In vendor benchmark tests he reported, end-of-day batch processing that once took roughly four hours completed in as little as eight seconds for a billion exposure records. When the platform is that fast, governance rather than technology sets the pace of what goes live. The industry's own customers appear to sense it. Sanjay Singla, chief executive of the trading-software vendor Commverse, says his firm's customer surveys consistently surface the same three concerns about agentic AI: cost, reliability and data security.

A rulebook rewritten around a hole

If adoption is running ahead of governance, the obvious question is what the regulators are doing about it. The answer, delivered in April 2026, was striking. The OCC, Federal Reserve and FDIC jointly rescinded and replaced their 2011 model-risk guidance, the framework known as SR 11-7 that most credit teams still reach for first. The new guidance states in its own text that generative and agentic AI models "are novel and rapidly evolving" and "are not within the scope of this guidance". The firm's broader risk-governance obligations, it confirms, still apply regardless.

Read that combination again. The rulebook was rewritten in 2026, and the exclusion was kept. That makes it a deliberate choice, not an oversight. Firms are required to govern these systems today, under existing model-risk, resilience and reporting law, and no regulator has written the standard that says how.

The pattern holds outside the US. The Bank of England's model-risk principles for banks require validation independent of the model's owner. But they formally bind only banks with internal model approval for regulatory capital, a category that does not include energy and commodity trading houses. The EU's AI Act targets AI that assesses the creditworthiness of natural persons, and its high-risk obligations are deferred to December 2027 in any case. A model assessing corporate counterparties sits outside its scope as drafted.

Call it a standard-shaped vacuum. Every serious credit function is already inside it.

What it costs when judgement fails

The scale of the exposure is not hypothetical, and the most instructive recent case involves no AI at all. Through the second half of 2026 a Singapore-based iron ore trader has faced allegations, which it denies and which remain before the courts, that it raised finance against invoices for trades that had never taken place. The alleged mechanism is the part worth holding onto. Invoices that had already been paid were offered as collateral, supported by contracts said to be fabricated, and the discrepancy survived because nobody compared the document against the counterparty whose name was on it. When a lender eventually asked that counterparty whether the trades existed, it was told they did not. Two major trading houses stopped dealing with the firm within days, and one European bank has provisioned around EUR 200 million against its exposure.

The older precedent is settled and quantified. In 2023, Trafigura, one of the world's largest commodity traders, booked a \$577 million charge, stated as its estimated maximum loss exposure, after discovering systematic fraud in nickel cargo against which it had extended credit and payment. That was a due-diligence and counterparty-credit failure, not an AI one, and no agentic system has yet produced a comparable, publicly disclosed loss. But it is the order of magnitude a credit function in this sector is already exposed to when the stated basis for a decision turns out not to be the real one. An ungoverned reasoning layer inside an agent is one more channel for the same failure mode.

And once an agent's output enters a regulated calculation or filing, the stakes sharpen further. Where an agent computes or feeds exposure numbers, netting assignments or a regulated report field, a hallucinated entry is not a display defect. It is a reporting breach.

Testing the reasoning behind the answer

The vendors are not blind to any of this, and the better products already constrain what their agents can do. The standard control is a fixed action vocabulary: the agent selects from a defined, enumerated set of recommendations and cannot invent a novel instruction. That is a necessary control. It is what makes the agent's behaviour testable at all.

It is not sufficient, because it bounds what the agent can do without bounding why it chose to do it. Return to the credit officer from the opening scene. Her agent was right about the exposure and wrong about why. That is a silent wrong-way-risk failure, the kind where the same price move that widens your exposure weakens the counterparty behind it. Only a check on the agent's reasoning, not its output, would have caught it before it mattered.

Testing the reasoning is a different discipline from reconciling a formula, but it is not mysterious. Run the same cases repeatedly and measure two things, not one: whether the agent reached the correct recommendation, and whether its stated reasoning was grounded in the right basis. Re-run a fixed benchmark on a schedule, and on any detected model-version change, to catch silent drift. Treat counterparty-supplied text, the financial narratives and filings the agent reads, as untrusted input, and test what happens when someone plants instructions in it.

Underneath all of it sits one design decision. The consequential numbers, exposure, limit utilisation, collateral haircuts, should be computed by deterministic code that the AI can annotate but never overwrite. Get that right and everything else is testable. Get it wrong and no amount of audit-trail discipline recovers it.

How much autonomy is the question, not how much capability

The constraint on autonomy in credit risk is not what agents can do. It is what a firm should permit them to do, and the gap between the two is a governance decision that has to be made deliberately rather than inherited from a vendor default. It helps to treat autonomy as a ladder rather than a switch.

AGENTIC ASSURANCE

How much authority should an agent hold?

Permitted operating authority, which is a governance decision rather than a technical ceiling.



Height denotes position on the scale, not preference. A deployment may sit at different levels in different domains.

Figure 2 · Six levels of permitted authority. Level 0 keeps existing model and control governance. Level 1 needs accuracy testing and training against over-reliance. Level 2 needs groundedness testing and read-only access. Level 3 needs a fixed action vocabulary, approval gates and full traceability. Level 4 needs enumerated write rights and sampling review.

Most credit work belongs at Levels 2 and 3: agents that identify, investigate, propose and evidence, with humans acting. Level 4 is defensible only where a wrong action is bounded and recoverable, such as re-requesting a missing confirmation. Level 5 sits badly with credit decisions and regulated filings, where accountability is personal and statutory. Within any level, oversight should be tiered by consequence. Small, collateralised exposures can reasonably be agent-adjudicated with sampling review. Large, unsecured or wrong-way names must route to an officer, with the agent as adviser

rather than decider. And when officers start overriding the agent in clusters, on a particular counterparty type or on the agent generally, that pattern is itself a signal. It should trigger a review, not sit in a log.

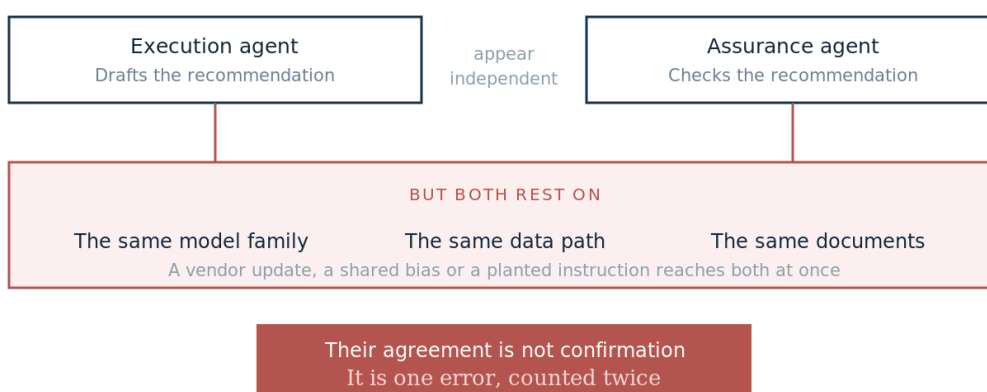
One error, counted twice

Keeping a human on every action, as Levels 2 and 3 do, is not the same as keeping a human across every pattern. One officer reviewing one recommendation cannot see drift accumulating across the thousand behind it, and that work runs at the agents' tempo rather than the reviewer's. The natural answer is to appoint another agent to do the watching. That instinct is right. What it hides is that two agents can agree with each other for reasons that have nothing to do with either being correct.

If the executing agent and the checking agent run on the same underlying model, draw on the same data pipelines, or read the same poisoned document, then their agreement is not confirmation. It is one error counted twice. The failure modes are specific, and they compound. Both agents can share a common model failure, or hallucinate in the same direction, or depend on the same upstream data. A prompt manipulation planted in a single document can propagate across both, and a vendor update can drift them in step. Agents that look separate on an architecture diagram can inherit the same context and lose the independence they appear to have.

THE INDEPENDENCE PROBLEM

One error, counted twice



WHAT GENUINE INDEPENDENCE REQUIRES

- Different model families
- Separate data paths and credentials
- A deterministic cross-check in every chain
- An opinion from outside the estate

Figure 3 · Two agents can be separately configured and still fail together, because independence is a property of what sits beneath them.

There are real mitigations. Use different model families for execution and assurance rather than different prompts to the same one. Separate the data paths, the credentials and the evidence stores. Put a deterministic cross-check at the boundaries, so that at least one link in every chain does not reason at all. And hold the assurance agent to the standard it exists to enforce: a control is not effective because an agent reports that it passed, so every conclusion it reaches has to carry evidence a person can inspect without re-querying the agent that produced it.

What those mitigations reduce, they do not eliminate. Model diversity is bounded by how few frontier model families exist and how much training data they share. Architectural independence is asserted at design time and erodes through exactly the convenience integrations that erode every other boundary. And each of those safeguards is configured, operated and reported on by the same firm that owns the estate. A firm can build a genuinely good assurance layer and still be unable to demonstrate from the inside that it is working. The reason has nothing to do with the care its people take, and everything to do with the fact that the assessment and the thing assessed share an owner. That is why no firm audits itself.

Start with the inventory

The uncomfortable finding, in most firms, comes before any of that testing begins. Agentic tools frequently sit outside the model inventory entirely. Nobody registered them, because nobody was sure

they were models. The first question for any credit function is therefore the simplest one: where does agentic or generative AI already touch our credit process, proposed or live?

The second question is who checks the answer, and it does not have a comfortable internal solution. A vendor cannot issue an independent opinion on its own product, and a firm that resells the AI cannot credibly critique it either. This is the same logic that pushed regulators to separate audit from consulting. Independence and scale sit in tension once the same firm sells both the transformation and the assurance opinion.

The strongest objection to an outside check comes from a capable model-risk team, and it deserves stating at full strength. Independent model validation is already a standard second-line function in most large banks, genuinely separated from the people who build and own the models, and where that separation exists a firm can apply this discipline to its own agents internally. Buying it from outside would add cost without adding independence. The objection is correct on its own terms. It describes banks, though, rather than the energy and commodity trading houses this is arriving at, and most of those firms have never needed a standalone model-validation function, because until agentic AI turned up almost everything they modelled was deterministic and was validated the old way, by reproduction. Testing reasoning quality under non-deterministic conditions is a different discipline from reproducing an exposure calculation, and standing it up from scratch to answer one question about one system is a poor use of a credit function's time.

Whoever provides that check, their opinion is evidence for the firm's own governance to weigh, not a warranty that replaces it. The decision to rely on, restrict or retire an agent stays with the credit function and its board, exactly as it would after any other independent technical opinion.

The credit officer in the opening scene will be fine this time, because the rating and the collateral happened to agree. The firms that remain fine will be those that know, before the next cold snap, which of their agents' answers rest on reasons that hold. They will also be able to demonstrate it to someone who does not work for them. That work begins with the inventory, because every other question in this argument depends on knowing what is already running.

The questions that follow the inventory are harder, and they are where most of the work sits. Which domains should be assured first, and against which authoritative source? How much authority should an agent hold in each one, and what has to be true before it is granted? What evidence must a conclusion carry before a credit officer is entitled to rely on it? And what should an independent opinion contain before a board treats it as evidence rather than reassurance?

The working paper

Those questions are the subject of the working paper behind this article, "When the Credit Officer Is an Agent: agentic assurance in energy and commodities trading," by Michał Falkowski, PhD. It sets out:

- ten assurance domains, each mapped to the risk it controls and the person who remains accountable for it
- the six-level autonomy model summarised above, with the governance every level demands
- a four-layer target architecture built around a deterministic core, and the rule that keeps the boundary intact
- a control matrix specifying permitted authority and required evidence, domain by domain
- ten failure modes of the assurance layer itself, each with the governance response that answers it
- the anatomy of a signed independent assurance opinion, down to the confidence interval and the named partner who stands behind it

The paper is available on request at horizoncreditadvisory.com.

Seamus Maher is a partner at Horizon Credit Advisory, a specialist counterparty credit risk advisory for energy and commodity trading. The firm works with credit, risk and IT teams building, replacing or reviewing the models and systems behind their credit decisions. It sells no agent and takes no vendor economics on any system it assures. The autonomy model and the assurance framework described in this article are drawn from the working paper cited above.

The opening scenario is illustrative, not client data. The credit officer is a composite; the failure she is looking at is real. Regulatory positions cited are current to 10 August 2026. The autonomy model is a conceptual proposition, not a description of established industry practice. This article does not constitute legal or risk-management advice.