SUPPLY CHAIN RESEARCH · EDITORIAL FEATURE

The Agentic

Accountability Gap

Decision Authority Is Being Granted Faster Than Anyone Is Deciding Who Answers For It

Agents are being given authority to release orders, reroute shipments, and reallocate inventory. The liability is disclaimed upstream, the logs record actions rather than reasoning, and the reversal path has usually never been tested. In the reference case, running it made the loss larger.

Independent, vendor-neutral research for supply chain and technology leaders

Supply Chain Research | supplychainresearch.com | 2026

Contents

Decision Authority Is Being Granted Faster Than Anyone Is Deciding Who Answers For It

01 Executive summary

Autonomous agents are being granted decision authority in supply chain operations. Not recommendation authority, which organizations have been comfortable with for years, but the ability to take an action with commercial consequence: to release a purchase order, reroute a shipment, reallocate constrained inventory between customers, or accept a supplier substitution. The deployment timeline for this authority is measured in quarters. The three structures that would make the authority accountable are being built on a timeline measured in years, and in most organizations have not been started.

Those three structures are liability allocation, a decision-level audit trail, and a tested reversal path. On the first, the commercial terms published by model providers, platform vendors, and application vendors disclaim responsibility for outcomes and cap exposure at trailing fees, with the indemnities that do exist directed at intellectual property rather than at decisions; one major vendor states the position plainly, that when an agent makes a decision the customer's company owns the outcome. On the second, standard agent logging captures prompts, tool calls, and actions, which is a record of what happened rather than an explanation of why, and stochastic systems cannot reliably be rerun to reproduce a decision. On the third, the reference case in this field is a trading firm that lost approximately four hundred and sixty million dollars in about forty-five minutes, where warning messages had been generated and not read, and where the remediation applied under pressure made the position worse. We say honestly that the case for agent autonomy is strong: decision latency, decision volume beyond human capacity, and consistency are real advantages, and well-governed deployments exist. The argument here is that authority and accountability have come apart, and that the fix is cheap before an incident and unavailable after one.

USD 460m
lost in roughly 45 minutes in the reference case, where the attempted fix worsened the problem

7%
of global annual turnover, the maximum penalty under the European artificial intelligence regulation

23%
of surveyed supply chain organizations reported having a formal artificial intelligence strategy

The argument in brief

- **Authority is arriving faster than accountability.** Agents are being granted the ability to act on a deployment timeline of quarters, while liability allocation, audit trails, and reversal paths are on a timeline of years.
- **The liability has already been allocated, to you.** Every layer of the stack disclaims outcome liability and caps exposure at fees paid. The residual sits with the deploying organization by default rather than by negotiation.
- **Logs record actions, not reasoning.** Prompts, tool calls, and outcomes are captured. Which alternatives were considered, what was decisive, and what a rerun would produce generally are not.

- **The reversal path is the untested control.** In the reference case it existed, had never been executed, and applying it under pressure enlarged the loss. Most deployments cannot say whether theirs works.
- **Autonomy is not the problem.** Latency, volume, and consistency are real advantages and controlled autonomy is a legitimate strategy. Deploying ahead of the accountability structure is the problem.

02 Authority granted, accountability deferred

The distinction that organises this article is between a system that recommends and a system that acts, and it is a larger distinction than the incremental language around agent deployment suggests. Figure 1 sets out what is being granted against what has been resolved.

Figure 1. The gap is between what agents may do and who answers for it. Authority over consequential actions is being deployed on a quarterly cadence; the structures that would make it accountable are not.

A recommendation system produces an output that a person evaluates and either acts on or does not. Whatever the quality of the recommendation, the accountability structure is intact: a named human made the decision, the reasoning can be reconstructed by asking them, and the reversal path is whatever the organization already had for a human error. Organizations have decades of practice governing this arrangement, and most of their controls, approvals, and audit procedures assume it.

An agent with decision authority removes the human from the transaction. The action occurs, the commercial consequence follows, and the question of who decided has no clean answer: the model produced an output, the orchestration executed it, the configuration permitted it, and the person who approved the deployment approved a capability rather than this particular decision. Every control designed around a human decision point has been bypassed, not through any failure but because the decision point no longer exists in the form the control assumed.

This is not an argument that the transition is wrong. Section eleven sets out the case for autonomy at some length and it is a strong case. It is an argument that the transition changes what an accountability structure has to do, and that most organizations have deployed the capability without revisiting the structure. The evidence in the following sections is that the liability position has been settled by contract in the vendors' favour, that the logging produced is insufficient to reconstruct a decision, and that the reversal path is generally untested. Each of these is remediable. None of them is remediated by default.

A note on scope belongs here. This article is not about whether agents work, which is a capability question, nor about whether pilots reach production, which is an adoption question. It is about what happens when they do work, are in production, and take an action that turns out to be wrong. That event will happen, because every system that acts at volume eventually acts incorrectly, and the question is whether the organization has built the structures that determine what happens next.

There is an organizational dynamic that accelerates the gap and is worth naming. Agent capability is procured and deployed by technology functions, whose success is measured by whether the capability works. Outcome accountability sits with operations, finance, and risk, who are frequently informed of a

deployment rather than consulted on the authority it carries. The result is that the decision to grant authority is made by parties who do not bear its consequences, which is the classic condition under which risk is systematically underpriced.

The correction is procedural and inexpensive. An agent authority should require sign-off from the function that owns the outcomes before it goes live, with that function asked to state what it believes the worst plausible loss to be and whether it is willing to carry it. Organizations that institute this find that authorities get narrower and that the conversation surfaces the questions in this article naturally, because the person being asked to own an outcome asks how they would know it had gone wrong.

03 The reference case, and why it still applies

The best-documented case of an automated system with decision authority causing a large loss is not from supply chain, and the details are worth setting out precisely because the regulator's findings are unusually complete. Figure 2 sets out the sequence.

Figure 2. The reference case: automation without a tested way back. Sequence drawn from the securities regulator's administrative proceeding, in which roughly USD 460 million was lost in about 45 minutes.

The facts are these. A trading firm deployed new code to its order routing system across eight servers. A technician copied the code to seven of them and did not copy it to the eighth. The deployment repurposed a configuration flag that, on the un-updated server, reactivated a dormant function that had been out of use for years. Before the market opened, the firm's systems generated ninety-seven warning messages indicating that the dormant function was disabled on some systems and not others. The messages were not reviewed. When the market opened, the un-updated server began sending orders at a rate the firm had not anticipated.

What happened next is the part that matters most for agent governance. The firm identified that something was wrong and attempted a remediation, which was to remove the new code from the seven servers where it had been correctly installed. That action, taken under extreme time pressure by people reasoning about a system behaving unexpectedly, worsened the problem, because it removed the correct behaviour from the servers that had it and left the erroneous behaviour running. The total loss was approximately four hundred and sixty million dollars over roughly forty-five minutes, and the regulator imposed a penalty of twelve million dollars in the first enforcement action under its market access rule.

The regulator's own framing of the general lesson is worth quoting because it applies directly to agentic deployment: the speed with which automated systems enter orders into the marketplace can turn an otherwise manageable error into an extreme event with potentially widespread impact. The mechanism is not that automation makes errors more likely; it is that automation removes the interval during which a human would have noticed. A person entering orders manually would have observed the anomaly within a handful of transactions. The system executed millions.

Three features transfer to the supply chain agent context without modification. The first is that the failure was a deployment and configuration error rather than a flaw in the core logic, which is the most common failure mode in complex automated systems and the one least addressed by testing the model itself. The second is that warnings were generated and not read, which is what happens when alerting volume exceeds review capacity, a condition that agent deployments create by design. The third, and the most important, is that the reversal path existed, had never been executed under realistic

conditions, and made things worse when it was needed. An organization deploying agents should read this case as a description of its own risk rather than as a story about a trading firm.

A fourth feature of the reference case deserves attention because it is the one most likely to be dismissed as inapplicable. The failure originated in a deployment process, in the gap between what an engineer intended and what was actually running on each machine. Agent deployments involve considerably more moving parts than a single code release: model versions that providers update on their own schedules, prompt templates maintained separately from application code, tool definitions, retrieval indices, and policy configurations, any of which can drift out of alignment without a formal release event.

The practical implication is that an organization should be able to state, for any past moment, exactly what configuration was in force: which model version, which prompt, which tool set, which policy constraints. Very few can. That record is a precondition for reconstructing an incident and it is also the control that would have caught the reference case failure, since the discrepancy between servers would have been visible in a configuration audit. It is cheap to build at deployment time and impossible to reconstruct afterward.

04 Where the liability comes to rest

If an agent takes an action that causes a loss, the question of who bears that loss has in most cases already been answered, in documents the deploying organization signed without negotiating them. Figure 3 traces the position through the stack.

Figure 3. Where the liability for an autonomous decision comes to rest. Each layer disclaims or caps its exposure, the indemnities that exist are directed at intellectual property, and the residual sits with the deployer.

At the model layer, standard business terms provide services on an as-is basis, disclaim warranties including fitness for purpose, and cap aggregate liability at the fees paid over a trailing period, commonly twelve months. The indemnity offered, where one is offered, covers third-party intellectual property claims arising from the output rather than commercial losses arising from a decision. At the platform layer the pattern repeats, with copyright commitments that are conditional on the customer maintaining specified safety configurations and that again address intellectual property rather than operational consequence.

At the application layer the position has been stated with admirable clarity by at least one major vendor, whose published position is that when an agent makes a decision, the customer's company owns the outcome. This is not evasion; it is an accurate statement of where the risk sits under the contracts as written, and it is more useful to a buyer than the ambiguity offered elsewhere. Systems integrators carry professional liability insurance subject to its own exclusions, and their engagement terms generally limit exposure to fees for the specific engagement.

The cumulative effect is that the deploying organization holds the residual. This is not unusual in enterprise software and it is not in itself objectionable: the deployer chose the use case, set the parameters, and connected the agent to its systems, so there is a reasonable argument that the deployer is the right risk-bearer. What is objectionable is that most deployers have not read the terms with this question in mind, have not quantified the exposure, and have not established whether their insurance responds. The allocation has been made and only one side was paying attention when it happened.

The practical response is not to demand indemnities that no vendor will give. It is to read the terms before deployment with the specific question of what happens if the agent takes an incorrect action at scale, to quantify the plausible exposure by reference to the authority the agent has been granted, and to make an explicit decision about whether that exposure is acceptable. Where it is not, the correct lever is usually to bound the authority rather than to renegotiate the contract, which is the subject of

section fourteen. An organization that has done this has made a decision; one that has not has made the same decision by default and without knowing it.

A related question that deploying organizations rarely ask concerns the chain of responsibility when an agent acts on data supplied by a third party. Agents in supply chain settings routinely consume carrier tracking feeds, market price references, supplier confirmations, and forecast signals, any of which may be wrong. If an agent takes a costly action because an upstream data feed was erroneous, the contractual position with that data supplier is generally weaker still than the position with the software vendors, since data is commonly supplied with express disclaimers as to accuracy.

This matters because it changes where the diligence should be focused. An organization that has scrutinised its model provider terms and not its data supply terms has examined the layer least likely to be the proximate cause. The practical step is to enumerate, for each agent authority, which external data sources can materially influence the decision, and to establish what recourse exists if one of them is wrong. In most cases the answer is that there is none, which is a manageable position once known and a surprising one to discover during an incident.

05 The insurance market has started answering

A useful signal about where a risk is heading comes from the parties whose business is pricing it, and the insurance market has begun to respond to autonomous systems in a way that deploying organizations should be tracking.

The principal industry body that develops standardised policy forms filed an optional exclusion addressing generative artificial intelligence in mid-2025, with attachment to general liability renewals from the beginning of 2026, and has indicated that it is evaluating further options including for agentic artificial intelligence specifically. Errors and omissions carriers have reportedly begun adding exclusions of their own. The direction is consistent: exposures arising from autonomous systems are being carved out of standard coverage rather than priced into it, at least for now.

Two cautions belong with this. First, several of the specific details circulating about endorsement codes and litigation volumes come from trade press rather than from the originating filings, and this article does not rely on them; the verifiable point is that a standard-form exclusion was filed and that carriers are actively considering the category. Second, an exclusion filed as optional is not the same as an exclusion applied, and coverage varies by carrier, jurisdiction, and negotiated wording. An organization should establish its own position by asking its broker rather than by reading market commentary.

The strategic reading is more useful than the coverage detail. Insurers are pricing specialists with no interest in the technology debate, and their behaviour indicates that they regard the loss distribution from autonomous decision-making as poorly understood and potentially large. That is an assessment worth weighing against vendor confidence, and it argues for the same practical conclusion this article reaches throughout: bound the authority so that the plausible loss stays within what the organization can absorb, because the mechanism that would otherwise transfer that loss is being withdrawn while the capability is being deployed.

There is an adjacent legal development worth tracking, which concerns liability for decisions made by algorithms shared across competing firms. Litigation over algorithmic pricing has tested whether firms using a common software provider's pricing recommendations can be liable for coordination, and courts have been building doctrine in this area over recent years with mixed outcomes at the pleading stage. The relevance to supply chain settings is that agents in this domain frequently run on shared platforms trained on or informed by industry-wide data.

An organization should therefore consider not only what its agent decides but what it decides in common with competitors using the same platform. Where an agent's pricing, allocation, or capacity decisions are materially shaped by a shared provider's logic, the organization has acquired an exposure that is not addressed anywhere in its technology governance and that sits with competition counsel rather than with the risk committee. Raising the question early is inexpensive; discovering it through an inquiry is not.

06 Adoption is compounding, governance is flat

The scale of the gap can be seen by putting the projected trajectory of agent deployment alongside the measured state of governance in the same population. Figure 4 does this, with the provenance caveat that the forecast is vendor-adjacent and the survey figures are self-reported.

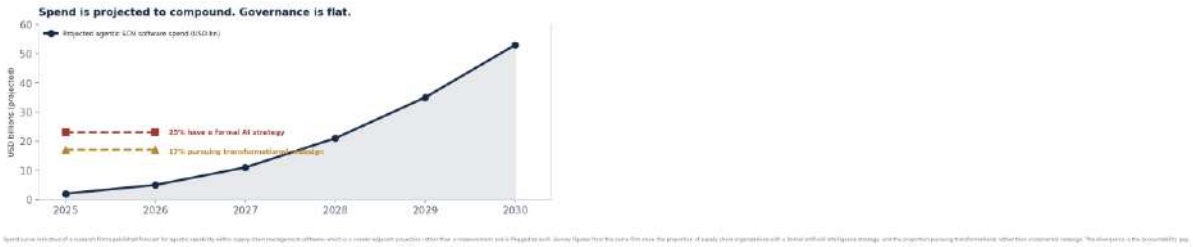

Figure 4. Spend is projected to compound while governance is flat. The forecast is a research firm's vendor-adjacent projection rather than a measurement; the survey figures are self-reported by supply chain organizations.

A widely-followed research firm projects that agentic capability within supply chain management software grows from under two billion dollars in 2025 to more than fifty billion by 2030, and separately that the proportion of enterprises using such software with agentic features adopted rises from around five percent to sixty percent over the same period. These are forecasts published by a firm that sells research to both vendors and buyers in this market, so they are directional rather than evidential, and they should be read as an indication of expected direction rather than as a measurement of anything.

Set against those projections are survey findings from the same firm that are measurements of a sort, being self-reported responses from supply chain organizations. Roughly twenty-three percent reported having a formal artificial intelligence strategy. Around seventeen percent reported pursuing transformational redesign of their operating model, with the substantial majority pursuing incremental application. The same firm has projected that more than forty percent of agentic artificial intelligence projects will be cancelled before the end of 2027, citing cost, technical debt, and inadequate risk controls among the reasons.

The divergence between these two sets of figures is the accountability gap expressed numerically. Deployment is expected to compound at a rate that implies most organizations in this population will be running agents with real authority within a few years. The governance indicators are flat and low. And the same source projecting the growth is projecting a high cancellation rate attributed partly to inadequate risk controls, which is an unusually direct statement that the governance deficit is expected to destroy a substantial fraction of the investment.

A deploying organization should treat the survey figures as a benchmark question rather than as an industry fact. If roughly a quarter of comparable organizations report a formal strategy, the useful exercise is to establish whether this organization is in that quarter, and specifically whether it can produce a document that states which decisions agents may take, who owns each authority, and what happens when one goes wrong. Most organizations that attempt this find they have deployment standards and no authority framework, which is the specific gap this article describes.

A separate signal in the same body of research concerns workforce composition, and it bears on accountability more directly than it first appears. A majority of surveyed supply chain leaders expect agentic capability to reduce entry-level hiring. Entry-level roles are where the routine exception handling currently sits, which means they are also where the tacit knowledge of what a normal decision looks like is acquired. An organization that removes those roles removes the population from which its future reviewers of agent decisions would have come.

This is a slow-acting risk rather than an immediate one and it deserves a place in the governance conversation nonetheless. The people who can tell whether an agent's expedite decision is sensible are people who have made that decision themselves several hundred times. If the pathway that produced them is closed, the organization retains the authority to override without retaining the capability to know when it should, which is the human oversight problem in its most durable form.

07 An action log is not a decision record

The second structural gap concerns reconstruction. If an agent takes a consequential action and the organization needs to establish afterward why, the logs it holds are generally insufficient, and the insufficiency is structural rather than a configuration oversight. Figure 5 sets out the difference.

Figure 5. An audit trail of actions is not an audit trail of decisions. Standard logging captures the observable transaction; reconstruction requires the alternatives, the weighting, and the decisive inputs.

What a standard agent deployment records is the observable transaction: the prompt or triggering event, the sequence of tool calls with their parameters, the responses received, the timestamps, the identity under which the agent operated, and the final action taken. This is a complete record of what happened and it is plainly useful for operational debugging, for establishing sequence, and for demonstrating that a particular action occurred at a particular time.

What it does not record is the decision. To explain why an agent chose to expedite one order and not another, an investigator needs to know which alternatives were available and considered, what weighting was applied to the competing objectives, and which specific inputs were decisive. None of these is emitted by default, and in most architectures none is recoverable after the fact. The intermediate reasoning, where it exists at all in an inspectable form, is typically not persisted, and where it is persisted it is a generated narrative rather than a faithful account of the computation.

The reproducibility problem compounds this. In a deterministic system, an investigator can rerun the inputs and observe the same output, which supports a reconstruction. Language-model-based agents are stochastic, and models are updated by their providers on schedules the deployer does not control, so rerunning the same prompt some weeks later may produce a different action for reasons entirely unrelated to the original event. The natural investigative method is therefore unavailable, and an organization that assumed it could simply rerun the scenario will discover this during the investigation rather than before it.

The practical remedy is to instrument for reconstruction deliberately, at the point of design. This means persisting the candidate set the agent considered, the ranking or scoring applied where the architecture produces one, the specific data records retrieved and used, the model version and configuration in force, and the policy constraints active at the time. This is more logging than most deployments produce and it is not technically difficult; it is simply not the default, and adding it after an incident recovers nothing. An organization that can produce this record can explain itself to a regulator, an auditor, an insurer, or a customer. One that holds only action logs can establish what happened and not why.

There is a distinction worth drawing between two kinds of explanation that organizations conflate when they discuss agent transparency. The first is a narrative account produced by the system itself, in which the agent is asked to describe its reasoning and generates a plausible-sounding explanation. The second is a record of the actual computation, comprising the inputs retrieved, the candidates evaluated, and the scoring applied. Only the second supports an investigation, and the first can be actively misleading because a generated rationalisation is not constrained to correspond to what actually drove the output.

This matters because generated explanations are readily available and satisfying, while computational records require deliberate engineering. An organization that has configured its agents to log their own reasoning has produced something that reads like an audit trail and does not function as one. Where the consequence is material, the record must be of what the system did rather than of what it says it did, and the difference should be understood by everyone who will rely on the logs.

08 What the regulation actually requires

There is one instrument that creates a hard obligation touching both logging and human oversight, and its scope and timing are worth stating precisely because a good deal of loose commentary surrounds it. Figure 6 sets out the position.

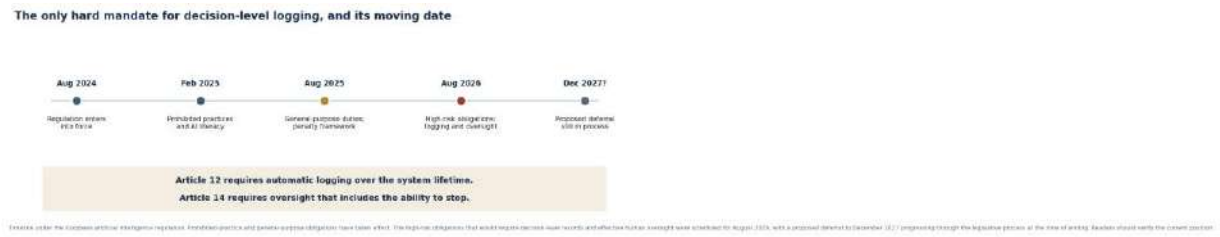

Figure 6. The only hard mandate for decision-level logging, and its moving date. High-risk obligations were scheduled for August 2026 with a proposed deferral to December 2027 progressing at the time of writing.

The European artificial intelligence regulation entered into force in August 2024 with a staged application timetable. Prohibited practices and literacy obligations applied from February 2025. Obligations on general-purpose models, the governance architecture, and the penalty framework applied from August 2025. The obligations attaching to high-risk systems, which are the ones relevant to autonomous operational decision-making, were scheduled for August 2026. A legislative package proposing to defer that date to December 2027 was progressing through the process at the time of writing, and organizations should verify the current position rather than relying on any secondary account including this one.

Two provisions matter most. The record-keeping provision requires that high-risk systems technically allow for the automatic recording of events over the lifetime of the system, at a level enabling traceability and supporting post-market monitoring. The human oversight provision requires that such systems be capable of being effectively overseen by natural persons, including the ability to interpret output, to decide not to use it, and to intervene or stop the system. Read together, these are the audit trail and the reversal path that this article has argued are missing, expressed as legal requirements.

The penalty structure is substantial: up to thirty-five million euros or seven percent of total worldwide annual turnover for prohibited practices, and up to fifteen million euros or three percent for breaches of the high-risk obligations. Those figures are what makes the classification question consequential, because they attach to the assessment of whether a given system falls within the high-risk categories, and most supply chain agent deployments have never been assessed against that classification at all.

There is an academic critique of the oversight provision that deploying organizations should read rather than dismiss. Analysis published in a peer-reviewed law and technology journal argues that the human oversight obligations are framed functionally rather than in terms of outcome, and do not straightforwardly vest an overseer with effective power to override. The practical version of this critique is familiar to anyone who has watched an operator monitor an automated system: nominal oversight where the human lacks the information, the time, or the standing to intervene is oversight in form only. An organization designing to satisfy the provision should design for the substance rather than for the audit.

The classification question deserves more attention than it usually receives, because it determines whether the obligations apply at all and most organizations have not performed the assessment. A supply chain agent that optimises routing is unlikely to fall within the high-risk categories. One that materially influences employment decisions, access to essential services, or safety components of regulated products may. The assessment is not difficult and it produces a documented position, which is worth having whether the answer is yes or no.

Where the assessment concludes that a system is out of scope, the obligations in the regulation remain a reasonable design reference. An organization that instruments logging and oversight to the standard the regulation would require, for a system it has determined is not covered, has built controls it can point to in any forum and has removed the risk that a later reclassification or a change in the system's use leaves it exposed. The cost of designing to the standard is modest; the cost of retrofitting is not.

09 What the voluntary standards settle

Alongside the regulation sit two widely referenced governance instruments, and it is worth being precise about what each does, because both are frequently cited as though they resolved the accountability question. Figure 7 compares them.

Figure 7. Three governance instruments, one enforceable requirement. The voluntary framework and the management-system standard are useful and neither compels decision-level records.

The risk management framework published by the United States national standards institute in January 2023 organises artificial intelligence risk work around four functions covering governance, mapping, measurement, and management, and it names accountability and transparency as characteristics of trustworthy systems. It is explicitly voluntary and non-prescriptive: it tells an organization what to think about and does not specify any particular control. That design is deliberate and appropriate for a framework intended to apply across every use of the technology, and it means that an organization can adopt the framework in full and still have no decision-level logging.

The international management system standard published in late 2023 is the first of its kind for artificial intelligence and is certifiable, which gives it a property the framework lacks: an external auditor examines whether the management system operates as described. But it is a management-system standard in the established tradition, concerned with policy, roles, risk process, and continual improvement rather than with technical controls, and it is generic across applications. A certified organization has demonstrated that it manages artificial intelligence deliberately. It has not demonstrated that any particular agent decision can be reconstructed.

The gap this leaves is specific and worth naming. The voluntary framework is outcome-oriented and unenforceable. The management standard is auditable and generic. The regulation is enforceable and specific but applies only to systems classified as high risk, and its relevant obligations have a date that has moved. There is consequently no instrument that currently guarantees, for a typical supply chain agent deployment, that a decision can be reconstructed or that a reversal path exists and works. Organizations that have adopted the framework or achieved certification have done something useful and should not conclude that they have closed this gap.

A practical way to use the two voluntary instruments together is to treat the framework as the agenda and the management standard as the discipline. The framework's four functions provide a reasonable structure for the questions a governance body should be asking about each agent authority, and the management standard provides the mechanism by which those questions get asked on a schedule rather than once. Neither supplies the technical controls, which the organization must specify itself.

What an organization should avoid is treating adoption of either as evidence that the accountability question is closed. Certification demonstrates that a management system exists and operates; it does not demonstrate that a specific agent decision can be reconstructed or that a reversal path has been executed. Where a board or a customer asks the accountability question, the useful answer describes the controls and the test results rather than the certificate.

10 The reversal path nobody has run

Of the three missing structures, the reversal path is the one most often assumed to exist and least often verified, and it is the one whose failure is most likely to convert a manageable incident into a serious loss. Figure 8 frames the diagnostic.

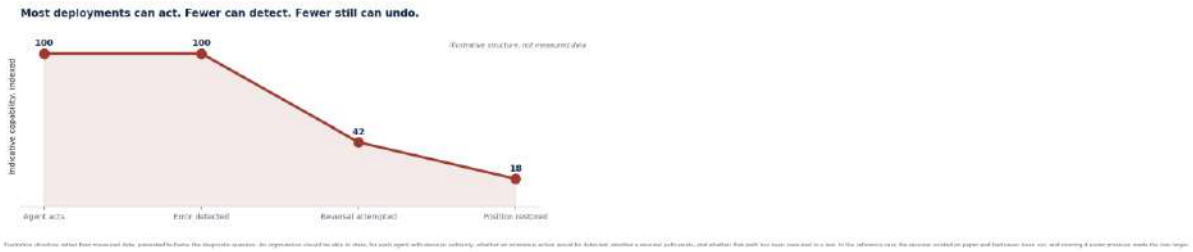

Figure 8. Illustrative structure, not measured data. Most deployments can act, fewer can detect an error promptly, and fewer still have executed a reversal under realistic conditions.

The question has three parts and an organization should be able to answer all three for every agent it has granted authority. Would an erroneous action be detected, and within what period? Does a defined path exist to reverse or contain the consequence? Has that path been executed, under realistic load and time pressure, by the people who would have to execute it? The reference case in section three failed the third test with a loss of four hundred and sixty million dollars, and the failure mode was specifically that the remediation, applied under pressure to a system behaving unexpectedly, was wrong.

Detection is harder in the supply chain context than in trading, and the difference cuts against the deploying organization. A trading error produces an immediate, visible, quantified position that the firm's own risk systems surface within minutes. A supply chain agent that has been systematically expediting the wrong orders, accepting a substitution it should have rejected, or reallocating inventory on a flawed priority rule may produce no signal at all for weeks, until the cost surfaces in a freight bill, a customer complaint, or a margin variance. The absence of a fast feedback loop means an agent can accumulate a large error before anyone knows to look.

Reversal in the physical world has properties that reversal in a database does not. An order released cannot be unreleased once the supplier has begun production; a shipment rerouted cannot be un-rerouted once the container is on the water; inventory allocated to one customer and consumed cannot be reallocated to another. The reversal path for many supply chain agent actions is therefore not a rollback at all but a mitigation, and the honest version of the question is what the organization would do to contain the consequence rather than to undo the action. Organizations that ask the question in the rollback form get a reassuring answer that does not describe their actual position.

The test that matters is an exercise, not a document. Select an agent with real authority, inject a plausible erroneous action into a controlled environment, and require the operations team to detect it and execute the containment on the clock. What such exercises typically reveal is that detection takes far longer than assumed, that the containment procedure references a system access that the person on duty does not have, and that the decision to invoke it requires an authority that is unavailable outside business hours. Each of those is trivially fixable once known and unfixable during an incident.

A further complication specific to supply chain settings is that the consequence of an erroneous agent action frequently lands on a third party before it lands on the organization. A customer receives the wrong allocation, a supplier receives a cancelled order after committing raw material, a carrier is booked and stood down. The organization may become aware of the error through the counterparty's complaint rather than through its own monitoring, which is both slow and expensive in relationship terms.

This argues for a specific and often-omitted control, which is to make the counterparty-facing consequences of agent decisions visible in the organization's own service monitoring. Where an agent has authority over allocation, the allocation exceptions should surface in the same review that a human allocator's exceptions would. Organizations that route agent decisions around their existing service governance because the agent is a technology deployment lose the detection mechanism they already had.

11 The fairness case: autonomy earns its place

This article has argued that accountability structures lag deployment, and a reader who concluded that agent autonomy is imprudent would be drawing a conclusion the evidence does not support and this article does not intend.

The strongest argument for autonomy is latency, and it is not a marginal one. Supply chain disruptions unfold on a timescale where the value of a response decays quickly: a port closure, a supplier failure, or a demand spike is worth responding to within minutes and considerably less worth responding to the following afternoon. A human decision process, with its queueing, escalation, and availability constraints, cannot operate at that speed across a large operation. An agent can, and the value of the faster response is real economic value that a slower and better-governed process forfeits.

The second argument is volume. A large operation generates far more decisions than its staff can attend to, so the practical alternative to agent decision-making is frequently not human decision-making but no decision at all: the exception sits in a queue, the default applies, and the opportunity passes. Where an agent handles the volume that would otherwise go unattended, comparing its decision quality to that of a careful analyst is the wrong comparison; the right comparison is against the default that would otherwise have applied, and against that benchmark agents frequently perform well.

The third argument is consistency. Human decision quality varies with fatigue, workload, time of day, and individual judgment, and that variance is itself a cost that is rarely measured. A system applying the same policy to every case eliminates that variance, which is valuable in its own right and is also what makes the decisions auditable in aggregate even where individual reconstruction is difficult. An organization can measure whether an agent's policy is producing good outcomes across thousands of decisions in a way it cannot for a dispersed human process.

A fourth point qualifies the article's framing rather than opposing it. The research firm whose adoption forecasts appear in Figure 4 also advises expanding autonomy in a controlled manner, beginning with low-risk decisions and building the data foundations and governance as the authority widens. That is a description of a responsible path and organizations are following it. The fair synthesis is that autonomy is not the risk; the sequence is. An organization that grants authority within bounds it has set, with reconstruction it has instrumented and a reversal it has tested, has done something both valuable and defensible. The gap this article describes is what happens when the first of those arrives without the other three.

A fifth consideration in fairness concerns the comparison base for reliability. Human supply chain decision-making has an error rate too, and it is rarely measured because errors are absorbed into normal variance and attributed to circumstance. An organization that holds an agent to a standard of near-perfection while never having quantified the error rate of the process it replaces has set an asymmetric test, and it will conclude that the agent is unreliable regardless of the evidence.

The disciplined approach is to establish the baseline before deployment. Sampling recent human decisions of the type in question, and assessing what proportion were suboptimal in hindsight, produces

a comparison that makes the agent evaluation meaningful. Organizations that do this generally find the human baseline considerably weaker than assumed, which both strengthens the case for automation and clarifies what standard the agent actually needs to meet.

12 Underwrite the authority, not the technology

The constructive principle follows from the diagnosis: govern the authority granted rather than the technology deployed, because the authority is what determines the exposure and it is the variable the organization actually controls.

This begins with an inventory that most organizations do not have. For each agent in production, the organization should be able to state what decisions it may take without human confirmation, what the value and volume limits are on those decisions, which systems it can write to, and who is the named individual accountable for the outcomes. That last element is the one most often missing. Technology ownership is always clear and outcome accountability frequently is not, and an authority without a named owner is an authority nobody is monitoring.

The second element is that the bounds should be enforced outside the agent rather than instructed within it. A limit expressed in a system prompt is a request; a limit enforced by the orchestration layer, the integration, or the downstream system is a control. This distinction is central and frequently collapsed in practice. An agent instructed not to approve orders above a threshold may still do so under an unusual input; an integration that rejects such orders regardless of what the agent requests cannot. Where the exposure is material, the bound belongs in the plumbing.

The third element is proportionality between authority and evidence. The authority granted should correspond to what the organization has actually established about the agent's performance in its own environment, on its own data, at its own volumes. Vendor benchmarks and pilot results in a controlled setting are weak evidence for production authority. A deliberate progression, in which an agent operates in recommendation mode while its decisions are compared against human ones, then takes low-value decisions under review, then takes them without review, generates the evidence that justifies each expansion. That progression is what controlled autonomy means, and it is materially different from deploying with full authority and monitoring for problems.

A fourth element concerns the aggregation of authorities, which organizations tend to evaluate individually and experience collectively. Several agents each holding modest authority over adjacent processes can, in combination, produce an outcome none of them was authorised to produce, particularly where the output of one becomes the input of another. A demand signal adjusted by one agent, consumed by a replenishment agent, and acted on by a procurement agent has passed through three modest authorities and generated a large commitment.

The governance response is to inventory authorities as a map rather than as a list, identifying where one agent's output feeds another's input, and to place the bound at the point where the cumulative commitment is made rather than at each individual step. This is the same principle that governs approval hierarchies for human decisions, where a chain of individually authorised steps that produces an unauthorised aggregate is a recognised control failure. The principle transfers directly and is rarely applied.

13 An agent authority protocol, and a scoring rubric

The principles combine into a protocol and a rubric that a risk committee or an operating committee can apply before granting an agent decision authority. Figure 9 sets out the discipline.

Figure 9. Underwriting an agent before granting it authority: name the accountable human, bound the authority outside the agent, log the decision rather than the action, test the reversal, and read the contracts.

The protocol runs as follows. Name an accountable individual for every agent authority, by name and not by function. Express the authority as explicit bounds on value, volume, and category, and enforce those bounds in the orchestration or integration layer rather than in instructions to the agent. Instrument for decision reconstruction at design time, persisting the candidate set, the decisive inputs, the model version, and the active policy constraints. Execute the reversal path as a timed exercise before the authority goes live and repeat it periodically. And read the vendor terms with the specific question of where an outcome loss lands, quantifying the exposure against the authority granted.

A scoring rubric

The dimensions below distinguish an underwritten authority from a deployed capability.

Dimension	Underwritten	Deployed only
Accountability	Named individual owns the outcomes	Technology owner, no outcome owner
Authority bounds	Enforced in orchestration or integration	Instructed in the system prompt
Audit trail	Alternatives, decisive inputs, model version	Prompts, tool calls, and outcomes
Reversal path	Executed as a timed exercise	Documented, never run
Detection	Stated detection window per authority	Assumed to be prompt
Liability position	Read, quantified, and accepted explicitly	Accepted by default in the terms
Basis for authority	Own performance evidence at own volumes	Vendor benchmark and pilot result

An authority scoring in the left column can be explained to a regulator, an insurer, an auditor, or a board after an incident. One scoring in the right column will be explained after the incident by whoever is available, from logs that record what happened and not why, under terms that place the loss with the organization. The rubric adds days rather than months to a deployment, and every element of it is cheaper before the authority is granted than after it has been exercised incorrectly.

14 Bounding authority so failure stays small

Of everything in the protocol, the bounding of authority deserves separate treatment because it is the control that determines the size of the loss when the other controls fail, and because it is the one an organization can implement immediately without any vendor cooperation.

The insight is that the maximum loss from an agent acting incorrectly is a function of the authority it holds, not of the probability that it errs. An organization cannot reliably estimate the error rate of a stochastic system operating on inputs it has not yet seen, and any figure it produces will be an extrapolation from a period during which the input distribution was different. It can, however, calculate exactly what the worst plausible outcome is if the agent behaves incorrectly at its authorised limit for the duration of its detection window. That calculation is arithmetic rather than forecasting, and it is the number a risk committee should be looking at.

Expressing bounds well requires more than a value cap. Useful dimensions include the value of an individual action, the cumulative value over a period, the number of actions per interval, the categories or counterparties in scope, the systems the agent may write to, and whether an action is reversible in the physical world. A cap of ten thousand dollars per order combined with unlimited order volume is not a bound. The cumulative and rate limits are frequently the ones that matter, because the failure mode that produces a large loss is generally repetition rather than magnitude, as the reference case demonstrates.

The enforcement point matters as much as the bound itself. Constraints expressed in the prompt are subject to the same variability as everything else the model produces, and an unusual input can produce an action outside them. Constraints enforced by the orchestration layer, the API gateway, or the receiving system apply regardless of what the agent generates. Where the exposure is material this distinction is the difference between a control and an aspiration, and an organization reviewing its deployments should ask specifically where each stated limit is actually implemented.

A final element is the relationship between bounds and detection. Bounds and detection windows multiply: an agent authorised to act at a given rate, with an error that would be caught within a stated period, produces a maximum exposure equal to the product. This gives the organization two levers rather than one and lets it choose the cheaper. Where detection is slow because the consequence surfaces in a downstream financial process, tightening the bound is the practical response. Where the bound is commercially constraining, investing in faster detection buys back the authority. Making that trade explicitly is what separates an underwritten authority from one that was set at whatever the pilot happened to use.

A final refinement concerns how bounds should change over time, which is a question most deployments never revisit. An authority set conservatively at launch on the basis of no operating evidence should be widened as evidence accumulates, and an authority that has produced a near miss should be narrowed. Neither adjustment happens by default, because the deployment is complete and attention has moved elsewhere, with the result that bounds set on the first week's information remain in force indefinitely.

Instituting a scheduled review of every agent authority, at a defined interval and with the performance record in front of the reviewers, converts a static configuration into a managed control. Organizations that do this find the review is short, that most authorities are widened on evidence, and that the occasional narrowing is precisely the intervention that prevents the incident. The value lies in the schedule rather than in any individual review.

15 Conclusion: someone owns the outcome

The transition from systems that recommend to systems that act is the most consequential change in supply chain technology in a decade, and it is being made without a corresponding change in how accountability is structured. That is the argument of this article, and the evidence for it is drawn from contracts that organizations have signed, logging architectures they have deployed, a regulatory timetable that has moved, and a regulator's account of what happened the last time an automated system with decision authority failed at speed.

The three gaps are specific. The liability has already been allocated to the deployer through terms that disclaim outcome responsibility and cap exposure at fees paid, while the insurance market that might otherwise absorb it is filing exclusions. The logging captures actions rather than decisions, and stochastic systems cannot be rerun to recover the difference, so an organization asked afterward why an agent did what it did will generally not be able to say. And the reversal path is the control most often assumed and least often tested, in a domain where many actions cannot be undone at all once a supplier has begun production or a container has sailed.

None of this argues against autonomy. The case for it is strong and this article has set it out at length: disruptions unfold faster than human decision processes can respond, the volume of decisions in a large operation exceeds what staff can attend to, and consistency is worth having. Organizations pursuing controlled autonomy, expanding authority as evidence accumulates, are doing something both valuable and defensible. The gap is what happens when authority arrives first and the rest is deferred, which is the common case rather than the exception.

What a supply chain leader should do is short and available now. Inventory every agent authority in production and name an accountable individual for each. Express the authority as explicit bounds on value, cumulative value, rate, and category, and move the enforcement of those bounds out of the prompt and into the orchestration or the receiving system. Instrument for decision reconstruction rather than action logging, persisting the candidate set, the decisive inputs, and the model version in force. Run the reversal path as a timed exercise before the authority goes live, because the reference case is a firm that had one, had never executed it, and made the loss larger by trying. And read the terms with the specific question of where an outcome loss lands, then decide explicitly whether the exposure is acceptable at the authority granted. The vendor has already told you the answer to the accountability question, in the plainest available language: when an agent makes a decision, your company owns the outcome. The only question left is whether the organization has arranged itself accordingly.

16 Methodology, caveats, and sources

Methodology

- This article draws on regulatory enforcement records, primary legislative texts, published commercial terms, standards documentation, peer-reviewed legal analysis, and industry research, current to mid-2026. Supply Chain Research is independent and accepts no payment from the model providers, software vendors, research firms, or insurers discussed.
- Market forecasts and adoption projections published by research firms that sell to both vendors and buyers in this market are identified as vendor-adjacent and are presented as projections rather than measurements.

Caveats

- The application date for the high-risk obligations under the European artificial intelligence regulation was subject to a proposed deferral progressing through the legislative process at the time of writing. Readers must verify the current position before relying on any date stated here.
- Commercial terms cited are drawn from publicly available standard agreements and may differ from negotiated enterprise terms. Organizations should read their own executed contracts rather than relying on this summary.
- Insurance market developments are drawn from trade reporting and filing summaries rather than from originating policy filings. Specific endorsement identifiers and litigation volume figures circulating in commentary were not verified and are not relied upon here.
- The reference case loss figure of approximately USD 460 million is the securities regulator's figure. The firm's own reported pre-tax loss was approximately USD 440 million. The regulator's figure is used as the primary source and the difference is noted.
- Figures 1, 3, 5, 7, and 9 are conceptual illustrations of structure. Figure 8 is explicitly illustrative and presents no measured data. Figure 4 combines a vendor-adjacent forecast with self-reported survey figures.
- This article addresses accountability structures for agents holding decision authority. It does not assess the capability of any model, product, or vendor, and does not constitute legal, insurance, or regulatory advice.

Sources

[1] US Securities and Exchange Commission. Administrative proceeding, Release No. 34-70694, In the Matter of Knight Capital Americas LLC (October 2013).

[2] European Union. Artificial Intelligence Act, Article 14: human oversight.

[3] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework, AI 100-1.

[4] International Organization for Standardization. ISO/IEC 42001:2023, artificial intelligence management systems.

[5] OpenAI. Business terms, including liability limitation and indemnity scope.

[6] Salesforce. Published guidance on governing artificial intelligence agents and outcome ownership.

[7] Gartner. Press release projecting agentic artificial intelligence adoption in supply chain management (vendor-adjacent forecast).

[8] Law, Innovation and Technology (Taylor and Francis). Peer-reviewed analysis of human oversight obligations under the European artificial intelligence regulation.

Additional context drawn from published Microsoft copyright commitment terms, from legal analyses of algorithmic pricing litigation, from insurance trade reporting on generative artificial intelligence exclusions, and from further research-firm survey publications which are identified as vendor-adjacent wherever used. This article is analysis, not legal, insurance, or regulatory advice, and its conclusions should be validated against your own contracts, policies, and advisers before any decision.

Supply Chain Research is an independent, vendor-neutral research platform for supply chain and technology leaders. We accept no payment from the model providers, software vendors, research firms, or insurers discussed. This article is analysis, not legal or insurance advice, and its conclusions should be validated against your own circumstances before any decision.