SUPPLY CHAIN RESEARCH · EDITORIAL FEATURE

AI Washing in Supply Chain Software

Separating Agentic Reality from Marketing

Every vendor now sells artificial intelligence, and much of it is real and valuable. But a large share is conventional software in new packaging, and regulators, analysts, and failed pilots are exposing the gap. This is how to tell shipped capability from slideware, and how to buy the real thing.

Independent, vendor-neutral research for supply chain and technology leaders
Supply Chain Research | supplychainresearch.com | 2026

Contents

A decision framework for the AI era

01 Executive summary

There has never been a worse time to take a vendor's word for what its software does, and never a more important one to check. Artificial intelligence is real, and in specific places across the supply chain it delivers substantial value today. But the label has become so commercially valuable that it is now attached to almost everything, and a large share of what is marketed as AI is competent conventional software in new packaging. The gap between the claim and the capability has grown wide enough that regulators have begun bringing enforcement actions, analysts have coined a term for the practice, and the majority of enterprise pilots built on inflated expectations are failing to reach production. For a supply chain or technology leader evaluating planning, warehouse, transportation, or visibility systems, the ability to tell shipped, production-grade capability from pilots and slideware is now a core procurement skill, not a technical curiosity.

This guide is written to build that skill, and it is deliberately even-handed. We are not skeptics of the technology. AI and machine learning materially improve demand forecasting, anomaly detection, route optimization, warehouse vision, and document processing right now, and the move toward agentic systems that can act, not merely recommend, is a real advance with real promise. But we say honestly that the marketing has run far ahead of the reality, that much of what is sold as intelligent or autonomous is neither, and that a buyer who cannot distinguish the two will pay premium prices for ordinary software and stake a program on capabilities that do not yet exist. The pages that follow define the practice of AI washing, trace the regulatory response, ground the reader in the technical spectrum from rules to true agents, map where AI is and is not real in supply chain in 2026, catalog the tactics vendors use to exaggerate, and supply the diligence questions, the proof-of-concept design, and the contract terms that separate the real from the marketed.

~130
of the thousands of self-described agentic AI vendors that Gartner judges to be real

40%+
of agentic AI projects Gartner expects to be canceled by the end of 2027

~95%
of enterprise generative AI pilots that MIT found delivered no measurable profit impact

What the evidence says

- **The label outran the capability.** A large share of products marketed as AI are conventional software relabeled, and even a supportive reading of the market finds only a small minority of self-described agentic vendors delivering real autonomy.
- **Regulators are now enforcing.** The SEC brought its first AI-washing cases in March 2024, and the FTC launched Operation AI Comply in September 2024, signaling that inflated AI claims carry legal risk.
- **Agent washing is the newest form.** Analysts describe widespread rebranding of chatbots, robotic process automation, and assistants as agentic AI, and predict most agentic projects will be canceled by 2027.

- **Pilots stall at the production line.** The majority of enterprise generative AI pilots fail to reach production or deliver measurable value, most often because they are brittle in real workflows rather than in demonstrations.
- **Diligence is the defense.** Buyers who ask what the AI actually does, demand production evidence on their own data, and tie payment to measured performance find the real value and avoid paying for the hype.

02 What AI washing is, and why it is everywhere now

AI washing is the practice of overstating the role, sophistication, or maturity of artificial intelligence in a product or service. The term is a deliberate echo of greenwashing, the older practice of overstating environmental credentials, and the parallel is apt: in both cases a label with genuine value and genuine meaning is applied loosely enough that it stops reliably signaling anything. A product is AI-washed when it is marketed as intelligent but runs on fixed rules, when a conversational interface is presented as autonomy, when a pilot is described as though it were in production, or when a capability on a roadmap is sold as though it were shipped. The common thread is a claim that the software is more intelligent, more autonomous, or more proven than it actually is.

The reason the practice is epidemic right now is straightforward: the label has become extraordinarily valuable, and the cost of applying it loosely has, until recently, been low. Investors reward AI, buyers seek it, and boards demand it, which creates enormous commercial pressure to describe whatever a product does in the language of artificial intelligence. The evidence that the label has outrun the substance comes from several directions at once, shown in Figure 1. An oft-cited early study by the venture firm MMC Ventures found that a large share of European startups classified as AI companies showed no material evidence of AI in their products. More recently, analysts examining the wave of agentic AI vendors concluded that only a small fraction of the thousands claiming the capability actually deliver it. And studies of enterprise adoption find that the overwhelming majority of generative AI pilots deliver no measurable financial impact. These are different populations measured in different years, and they should not be added together, but they point in the same direction: the gap between what is called AI and what functions as AI is large.

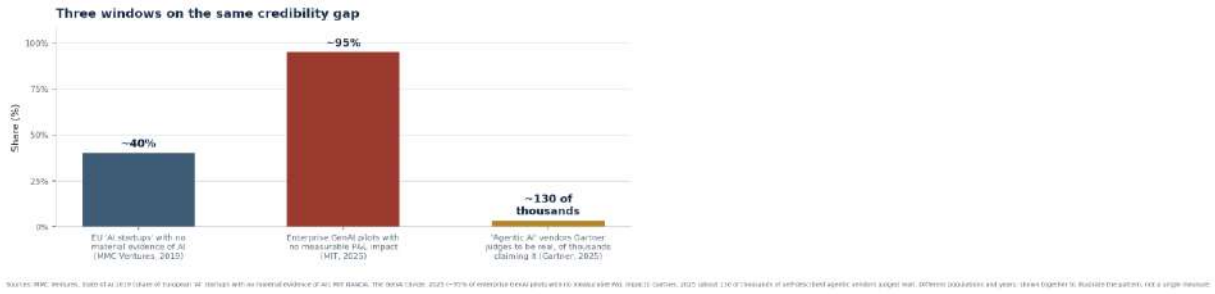

Figure 1. Three separate findings, from different years and populations, that each illustrate the distance between the AI label and the underlying capability.

For a buyer, the significance is not that AI is fake, which it is not, but that the label has lost its power to discriminate. When almost every product in a category claims to be AI-powered, the claim conveys no information, and the burden shifts to the buyer to determine what each product actually does. That determination is the subject of this guide. It begins with understanding that AI washing is not usually outright fraud, though it can be; more often it is a matter of degree and emphasis, of describing a rules engine as cognitive or a recommendation as a decision, and the ordinariness of the exaggeration is precisely what makes it hard to see and easy to pay for.

The greenwashing parallel is worth pressing, because it predicts how the AI label decays. When environmental claims proliferated without standards, the honest and the dishonest became hard for

buyers to tell apart, and the label's value collapsed until independent verification and clearer definitions restored some meaning. The AI label is on the same trajectory. As every product in a category claims intelligence, the claim stops distinguishing the products, and the buyers who continue to reward it simply pay a premium for a word. The eventual correction, as with environmental claims, will come from verification and definition, from buyers who demand proof and from a market that learns to price the label only when it is substantiated. This guide is an attempt to bring that correction forward for the individual buyer, who need not wait for the market to mature before demanding the evidence that separates the real from the labeled.

It helps to distinguish the forms AI washing takes, because they call for different scrutiny. The first is capability inflation, describing conventional software, rules, or analytics as artificial intelligence. The second is autonomy inflation, the agent washing described later, presenting an assistant or an automation as an autonomous agent. The third is maturity inflation, presenting a pilot, a preview, or a roadmap item as a shipped and proven capability. The fourth is provenance inflation, claiming as proprietary the capability of an underlying foundation model the vendor merely calls. A single product can exhibit several at once, and each is exposed by a different question: what the system does, how autonomously it does it, whether it is in production, and whose technology is actually at work. Naming the forms turns a vague unease about a claim into a specific line of inquiry.

03 The regulators have arrived

The clearest signal that AI washing has crossed from marketing excess into material risk is that regulators have begun to act, shown in Figure 2. What was recently treated as ordinary promotional enthusiasm is now, in some contexts, an enforceable misrepresentation, and the enforcement has come quickly and from more than one agency.

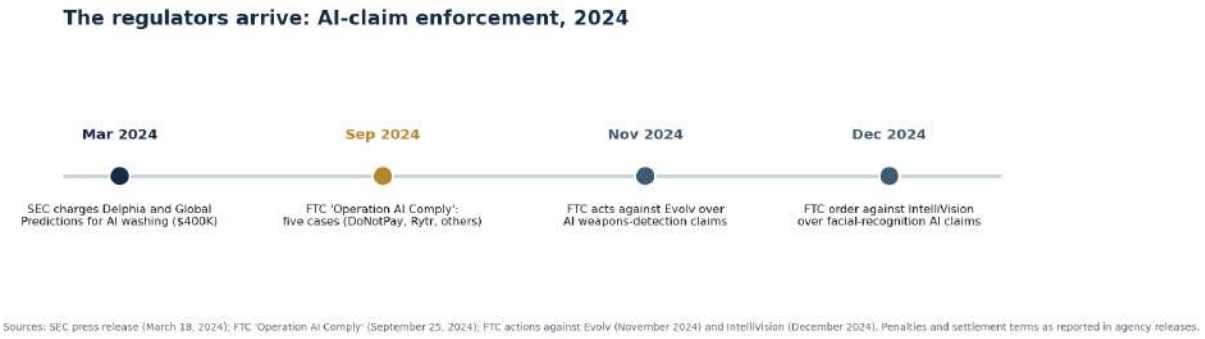

Figure 2. A compressed timeline of AI-claim enforcement in 2024. The pace and breadth signal that inflated AI claims now carry legal and financial risk.

The SEC's first AI-washing cases

In March 2024 the Securities and Exchange Commission announced its first enforcement actions specifically targeting AI washing, settling charges against two investment advisers, Delphia and Global Predictions, for making false and misleading statements about their use of artificial intelligence. The advisers agreed to pay civil penalties of two hundred twenty-five thousand and one hundred seventy-five thousand dollars respectively, a combined four hundred thousand dollars. The details are instructive because they define the line. Delphia had claimed to use client data to make its AI smarter and to predict which companies and trends would succeed, when, according to the Commission, it had not used the client data in that way and had not built the algorithm it described. The cases were brought under existing antifraud and marketing rules rather than any new AI-specific statute, which is the point the SEC was making: a false claim about AI is simply a false claim, and the agency did not need new authority to pursue it. The Commission's leadership warned publicly that firms should not, in its words, do the equivalent of AI washing, and the message to the market was that AI-related statements, wherever they are made, will be held to the same standard as any other material representation.

The FTC's Operation AI Comply

Six months later, in September 2024, the Federal Trade Commission launched what it called Operation AI Comply, a coordinated set of five enforcement actions against companies it alleged had used AI to supercharge deceptive or unfair conduct, or had simply lied about what their AI could do. The most prominent target, DoNotPay, had marketed itself as the world's first robot lawyer and claimed its service could substitute for human legal expertise; the Commission alleged the company had never tested its output against that of a lawyer and had employed none, and DoNotPay agreed to pay one hundred ninety-three thousand dollars and to warn customers about the limits of the service. Other cases in the sweep targeted business-opportunity schemes that used AI hype to lure investment, including one the

Commission said had cost consumers nearly sixteen million dollars, and a writing tool whose review-generation feature the Commission argued existed largely to produce fake reviews. The Commission followed with further actions, including one against a security vendor over its AI weapons-detection claims and another against a facial-recognition provider over its accuracy claims. Notably, one of the cases drew dissents from two commissioners who worried that penalizing a general-purpose tool for possible misuse could chill innovation, a reminder that the boundary between legitimate marketing and deceptive claim is truly contested at the edges.

For a supply chain buyer, the regulatory turn matters less for the specific cases, which involve consumer products, than for what it establishes. Inflated AI claims are now a documented enforcement priority, which means they carry disclosure and liability risk for the vendors that make them and, by extension, reputational and contractual risk for the buyers who rely on them. A vendor willing to overstate its AI to the market is a vendor whose other representations deserve scrutiny, and the regulatory record gives a buyer both a warning and a lever.

The dissents in one of the FTC's cases are worth dwelling on, because they mark the real difficulty at the boundary. Two commissioners objected to penalizing a general-purpose writing tool for the possibility that customers might misuse it, warning that holding a maker liable for a tool's potential misuse could chill legitimate innovation. The objection has force, and it points to a real tension: the same scrutiny that protects buyers from inflated claims can, applied too aggressively, punish honest builders of dual-use tools. The lesson for a buyer is not that every AI claim is fraudulent, which the dissents rightly resist, but that the line between confident marketing and deceptive claim is drawn by specifics, by whether a stated capability is real, proven, and as described, rather than by the mere use of ambitious language. That is the same line this guide asks buyers to draw, and it is why the diligence here targets evidence rather than enthusiasm.

04 Agent washing: the newest and most aggressive hype

If AI washing is the genus, its most virulent current species is agent washing, the rebranding of existing products as autonomous agents. The term was popularized by Gartner, which has documented a widespread pattern of vendors relabeling chatbots, robotic process automation, and simple assistants as agentic AI without adding the underlying capability that the word agent is supposed to denote. The commercial incentive is obvious: agentic AI is the most sought-after and most richly funded category in enterprise software, and attaching the label to an existing product is far faster than building a real agent. The result is a market in which the term has been diluted almost to meaninglessness, and in which buyers must work hard to find the minority of offerings that are real.

Gartner's assessment of that minority is striking. Of the thousands of vendors claiming agentic AI capability, the firm estimates that only about one hundred thirty are actually building something that deserves the label, with much of the rest amounting to chatbots, robotic process automation, and assistants in new packaging. The same analysis carries a prediction that has become the defining statistic of the moment: more than forty percent of agentic AI projects will be canceled by the end of 2027, undone by escalating costs, unclear business value, and inadequate risk controls. Figure 3 places that prediction against the investment posture that produced it. A poll of more than three thousand organizations found only a fifth had made significant investments in agentic AI, with far more making conservative bets or waiting to see, which suggests the market itself senses the gap between the promise and the present reality.

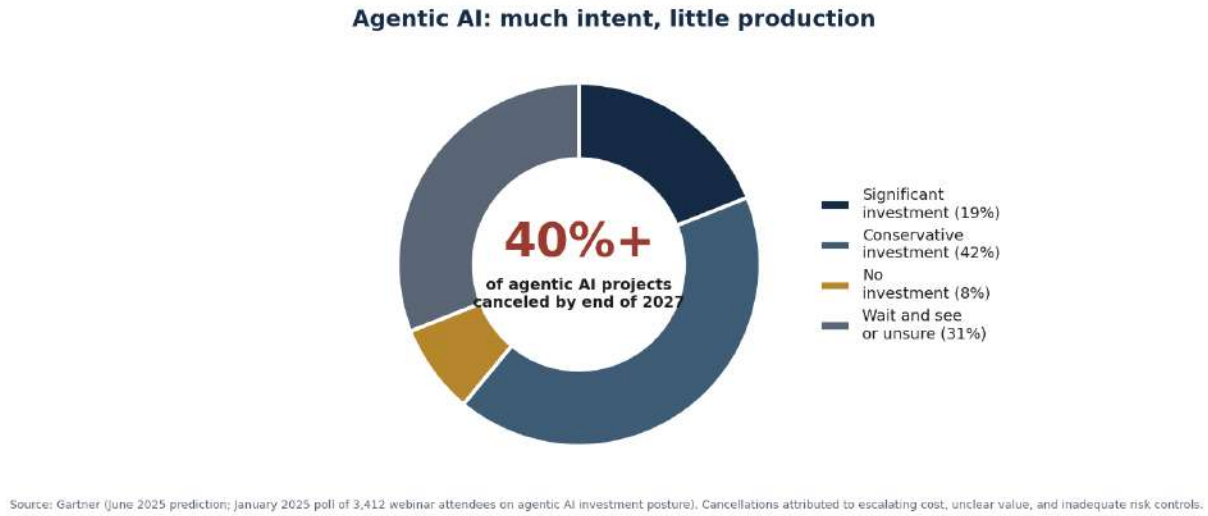

Figure 3. Investment posture toward agentic AI, and the cancellation Gartner expects. Most organizations are cautious, and most projects are predicted not to survive to 2027.

The gap is confirmed by a separate and telling divergence between enthusiasm and trust, shown in Figure 4. Research has found that while the overwhelming majority of organizations plan to increase their investment in agentic AI, only a small fraction trust AI agents to autonomously run core end-to-end business processes. That divergence is the honest center of the agentic moment: leaders are investing because the potential is real, and withholding trust because the present capability is not yet sufficient

for the autonomy the marketing implies. A buyer who holds both facts at once, investing in the potential while withholding trust from the claim, is reasoning correctly.

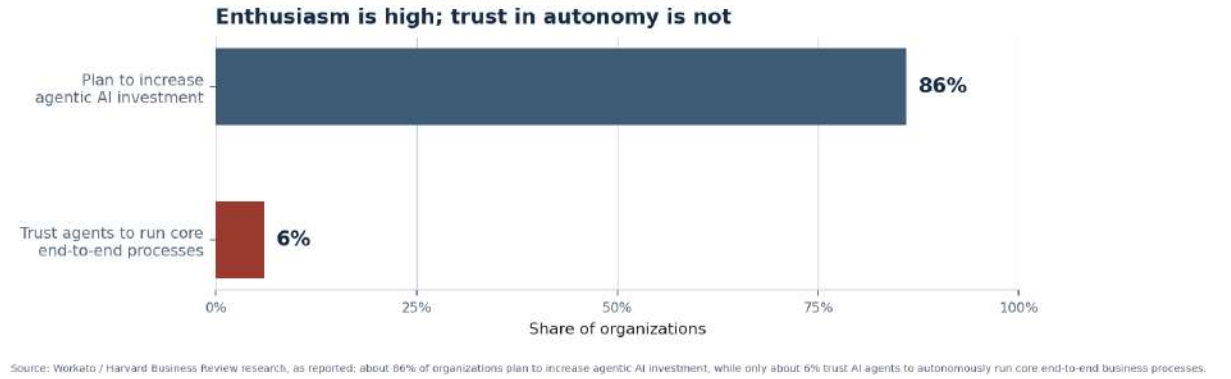

Figure 4. The distance between investment intent and operational trust. Enthusiasm for agentic AI is near-universal; willingness to let it run core processes autonomously is rare.

A concrete illustration makes the pattern recognizable. Consider a product that lets a planner type a question in plain language, receive an answer drawn from the planning system, and then execute the planner's chosen action with a click. Marketed plainly, this is a conversational assistant with a convenient action button, and a useful one. Marketed in the language of the moment, the same product becomes an autonomous agent that senses, decides, and acts, which implies an autonomy it does not have, because it answers when asked and acts only when a human chooses the action. Nothing about the product changed; only the label did. The buyer's task is to look past the label to the behavior, and to ask the question that settles it: what does this system do when no human is prompting or approving it. If the answer is nothing, it is an assistant, whatever the marketing calls it.

None of this means agentic AI is vaporware. It means the category is early, that the label has been applied far more widely than the capability warrants, and that the burden is on the buyer to determine, for any specific product, whether there is a real agent underneath. Doing that requires understanding what an agent actually is, which is where the technical grounding of the next two sections becomes indispensable.

05 The ladder from automation to agents

To see through AI washing, a buyer needs a mental model of the levels of capability that the single word AI is used to describe, because the exaggeration almost always consists of presenting a lower level as a higher one. The capabilities form a ladder, shown in Figure 5, and each rung is plainly useful, but the rungs are not the same, and the difference between them is precisely what a vendor obscures when it applies the most advanced label to a less advanced product.

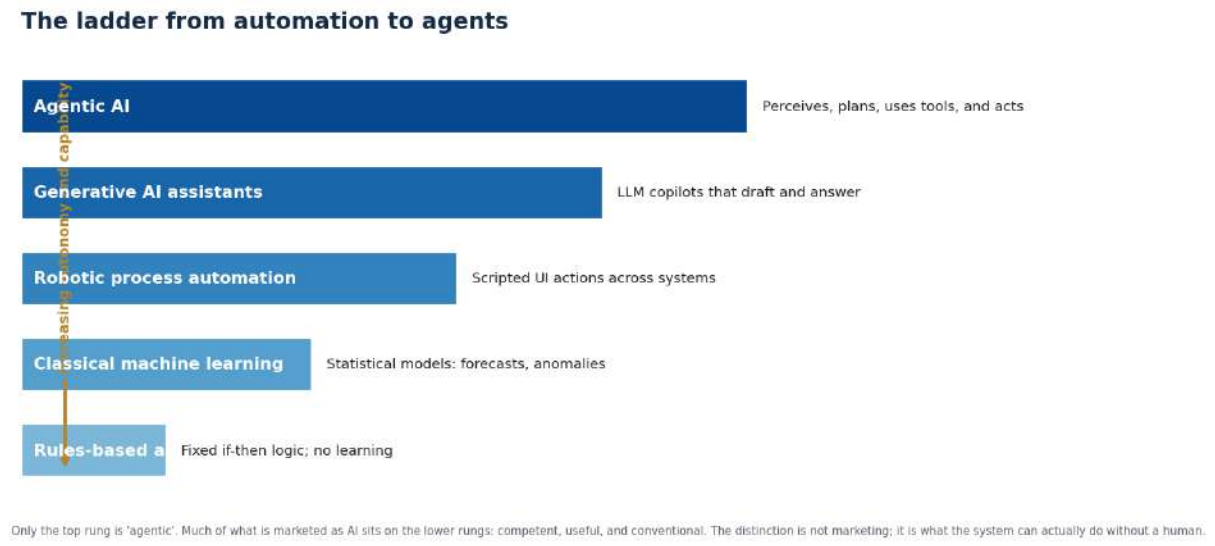

Figure 5. The ladder from fixed automation to autonomous agents. Each rung is useful, but only the top is agentic, and much of what is marketed as AI sits lower down.

The lowest rung is rules-based automation: fixed logic that follows predetermined instructions, valuable and reliable but incapable of learning or adapting. It is not AI in any meaningful sense, though it is frequently marketed as such. The second rung is classical machine learning, the statistical models that learn patterns from data to forecast demand, detect anomalies, or optimize a route. This is real AI, and in supply chain it is where most of the genuine value lives, but it is mature, well-understood technology rather than anything novel or autonomous. The third rung is robotic process automation, software that mimics the clicks and keystrokes a person would perform across systems; it automates work but follows scripts and does not reason, and relabeling it as an agent is one of the most common forms of agent washing.

The fourth rung is generative AI, the large language models that power the assistants and copilots that have defined the past few years. These systems can draft, summarize, answer, and converse with a fluency that is truly new, and they are a real advance, but a copilot that responds to prompts is an assistant, not an agent: it waits to be asked, and it recommends or produces rather than acting on its own. The fifth and highest rung is agentic AI, a system that perceives its environment, reasons toward a goal, plans a sequence of steps, uses tools, and takes action with some degree of autonomy. This is the frontier, it is real but early, and it is the label most likely to be applied to products that sit one, two, or three rungs below it. The practical value of the ladder is that it converts a vague claim, this product uses

AI, into a precise question, which rung is it actually on, and that question is the beginning of every honest evaluation.

The distinction between the rungs is not academic; it is commercial, because the rungs are priced differently. Software described as agentic AI commands a premium over software described as automation or analytics, even when the underlying capability is the same, because the label carries the promise of autonomy and the scarcity of the frontier. A buyer who cannot place a product on the ladder cannot tell whether the premium buys real capability or only the word, and vendors are aware of this. The ladder is therefore a pricing defense as much as a technical one: knowing that a product sits on the automation or assistant rung tells a buyer that its price should reflect automation or assistant value, whatever the marketing calls it, and equips the buyer to resist paying agentic prices for pre-agentic capability.

06 What agentic AI actually means

Because agentic is the most abused term in the market, it is worth defining with precision, so that a buyer can hold a specific claim against a specific standard. A true agent is distinguished from a chatbot or a recommendation engine by a combination of properties, and the absence of any one of them is usually the tell that a product is agent-washed.

- **Autonomy.** A real agent can operate without a human prompting each step, pursuing a goal across time rather than responding to a single request. A system that acts only when asked, and does only what it is asked, is an assistant, however sophisticated its language.
- **Goal-directedness.** An agent is given an objective, not an instruction, and works out how to achieve it. A chatbot is told what to say; an agent is told what to accomplish and determines the steps itself.
- **Planning and reasoning.** An agent decomposes a goal into a sequence of steps, adapts the sequence as conditions change, and reasons about trade-offs. Fixed workflows and decision trees, however elaborate, are not planning; they are pre-scripted paths.
- **Tool use.** An agent can call other systems, query data, run calculations, and invoke software to accomplish its goal, rather than being confined to generating text. The ability to act on the world through tools is central to what makes an agent an agent.
- **Memory and adaptation.** An agent retains context across steps and interactions and adjusts based on what it learns, rather than treating each request as isolated. The absence of memory is one reason many pilots fail, as later sections describe.
- **The ability to act, not merely recommend.** This is the decisive property. A system that surfaces a recommendation for a human to execute is a decision-support tool; a system that can take the action itself, within defined bounds, is an agent. Much of what is sold as agentic stops at the recommendation and leaves the human to act, which is valuable but is not autonomy.

The reason this precision matters is that vendors routinely satisfy one or two of these properties and claim the whole. A conversational interface on top of a recommendation engine has fluency but no autonomy. A scripted multi-step workflow has sequence but no planning. A system that acts only with a human approving every step has the appearance of agency but not its substance. The honest questions a buyer asks are therefore specific: does it pursue a goal or answer a prompt, does it plan or follow a script, can it act or only advise, and how much of the autonomy shown in the demonstration survives when a human is not watching. The gap between demonstration and production, where those questions are answered, is the subject of the next section.

A concrete contrast fixes the distinction. Suppose a planner needs to respond to a supplier delay. A conversational assistant, asked about the delay, retrieves the affected orders, summarizes the exposure, and, when the planner selects a mitigation, drafts the necessary messages. That is real and useful, and it is not an agent, because the planner drives every step. A true agent, given the goal of maintaining service through the delay, would itself identify the affected orders, evaluate mitigations against the plan's constraints, select one within its authorized bounds, execute the reallocation across the connected systems, and report what it did, returning to the planner only when the situation exceeds its

authority. The difference is not fluency, which both may share, but agency: the assistant waits and advises, while the agent pursues the goal and acts. When a vendor calls a product an agent, this is the behavior to look for, and its absence is the tell.

It helps to think of autonomy not as a switch but as a series of levels, because vendors often claim a higher level than they deliver. At the lowest level the system is purely advisory: it recommends, and a human decides and acts. Above that is supervised action, where the system proposes and a human approves each step before it executes. Higher still is bounded autonomy, where the system acts on its own within tightly defined limits and escalates anything outside them. At the top is broad autonomy, where the system pursues a goal across many steps with only occasional human oversight. Most supply chain capability marketed as agentic today sits at the advisory or supervised levels, which are useful and appropriate for high-stakes decisions, but which are not the broad autonomy the word agent tends to imply. A precise buyer asks which level a product actually occupies, and matches the level to the risk of the decision being automated.

07 Where AI is real in supply chain today

A guide this skeptical of AI claims owes the reader an equally clear account of where AI is real, because the purpose of seeing through the hype is not cynicism but discernment: to find and buy the substantial value that truly exists. In supply chain, that value is considerable, and it is concentrated in the areas where machine learning has matured over years of use, shown in Figure 6. The pattern is consistent: the most proven applications are classical machine learning on well-defined problems, while the least proven are the fully autonomous multi-agent systems at the frontier.

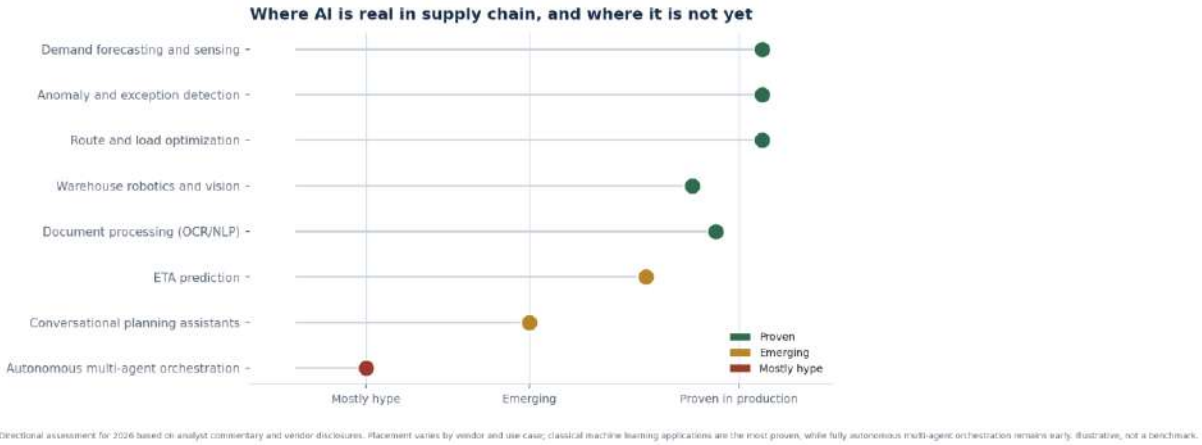

Figure 6. A directional map of AI maturity across supply chain use cases in 2026. The proven value is concentrated in mature machine learning; full autonomy remains early.

The proven ground

Several applications are real, in production, and delivering measurable value across many organizations. Demand forecasting and demand sensing use machine learning to improve prediction accuracy over traditional statistical methods, particularly for volatile or short-life products. Anomaly and exception detection surface the shipments, orders, and inventory positions that need attention out of volumes no human could monitor. Route and load optimization apply mature operations-research and machine-learning techniques to cut transportation cost and improve utilization. Warehouse robotics and computer vision guide picking, sortation, and inventory counting with a reliability that has made them standard in modern facilities. Document processing uses optical character recognition and natural-language processing to extract data from invoices, customs forms, and bills of lading, removing enormous manual effort. And estimated-time-of-arrival prediction, the core of modern visibility platforms, uses machine learning across historical and real-time data to forecast when freight will actually arrive. None of these is speculative; all are shipped, and a buyer should expect vendors to prove them on the buyer's own data.

The emerging frontier

Above the proven ground sits a band of capability that is real but earlier, and here the buyer must be more careful. Conversational assistants that let a planner query a system in natural language are shipping across the major platforms and are useful, though they are assistants rather than agents. The

leading planning and execution vendors have begun to introduce agentic capabilities: several now offer domain-specific agents for warehouse, logistics, and inventory decisions, and describe a progression from agents that monitor and analyze toward agents that take more autonomous action within a domain. These are genuine developments, and some are in customer hands, but the honest reading, which even vendor leaders sometimes offer, is that much of the fully autonomous, multi-agent orchestration remains in laboratories and early deployments rather than in broad production. One senior vendor executive candidly described end-to-end multi-agent operation as working in their labs but not yet widely present in customer deployments, which is exactly the kind of honesty a buyer should seek and reward. The frontier is worth engaging, but it should be bought as what it is, an early capability to pilot carefully, not a proven one to depend on.

The practical guidance that follows from this map is to match expectation to maturity. For the proven applications, a buyer should expect production references, measured results, and a proof of concept on their own data, and should treat any vendor who cannot provide them as suspect, because the capability is mature enough that evidence should be abundant. For the emerging applications, a buyer should expect candor about what is shipped versus piloted, a clear account of how much autonomy is real, and a design that keeps humans in the loop, and should be wary of any vendor who presents frontier capability with the confidence appropriate to proven capability. The map is not a reason to avoid the frontier; it is a guide to buying each capability with the right expectations.

The vendor landscape, read carefully

The major planning, execution, and visibility vendors are all investing heavily in AI, and several have introduced capabilities described as agentic. These are real developments worth engaging, but they span a wide range from shipped to announced, and the table below, drawn from public disclosures, is offered as a map to read carefully rather than a scorecard to trust. Every entry is a claim to verify directly against production references and a proof of concept, because the distinction between a capability that is live, one that is in limited availability, and one that is on a roadmap is exactly what a vendor's marketing tends to blur and exactly what a buyer must establish.

Vendor	Described capability (verify directly)	Maturity signal
Kinaxis (Maestro)	Maestro Agents; Agent Studio for no-code agents; orchestrator agents planned	Phased: some live, some limited, some roadmap
Blue Yonder	Warehouse, logistics, and inventory operations agents; end-to-end multi-agent	Some agents live; multi-agent largely in labs
o9 Solutions	Knowledge-graph model with agentic layers on top	Graph mature; agentic emerging
SAP	Joule assistant and business AI agents across the suite	Assistant shipping; agents expanding
Manhattan Associates	Agentic capabilities within the Active platform	Verify shipped versus announced
Visibility platforms	Machine-learning arrival prediction plus agents for exceptions	Prediction proven; agents emerging

Vendor capabilities and names are drawn from public disclosures and change rapidly; treat every entry as a claim to confirm rather than a fact to rely on.

The pattern across the landscape is consistent with this guide's central theme. The proven capability is machine learning applied to forecasting, optimization, and detection, which these vendors have shipped for years. The agentic capability is newer, is arriving in stages, and mixes shipped features with limited previews and roadmap ambitions, and the vendors that describe the distinction candidly, including one whose executive openly noted that end-to-end multi-agent operation works in their labs but is not yet widespread in customer deployments, are the ones a buyer can most readily trust. The table is a starting point for questions, not a substitute for the production evidence and the proof of concept on the buyer's own data that the later sections prescribe.

A word is owed on the return-on-investment figures that vendors attach to these capabilities, because even where the capability is real, the numbers require care. Vendor-reported returns are typically drawn from favorable deployments, measured by the vendor, and presented without the baseline against which the improvement should be judged. This does not make them false, but it makes them directional, and a buyer should treat a vendor's return figure as a hypothesis to test on its own data rather than a result to expect. The proven applications can deliver real and measurable returns, in forecast accuracy, inventory reduction, and labor productivity, but the return a specific organization will see depends on its own baseline, data quality, and processes, none of which the vendor's figure reflects. The dependable number is the one the buyer measures in its own proof of concept, not the one on the vendor's slide.

08 The gap between the demo and production

The single most important thing to understand about AI capability, and the thing AI washing most exploits, is that a demonstration and a production system are separated by a wide and treacherous gap. A demonstration is designed to succeed. It runs on curated data, follows the path the vendor intends, and shows the capability at its best. Production is the opposite: messy data, adversarial edge cases, unexpected inputs, integration with systems that behave in undocumented ways, and the requirement to work reliably thousands of times rather than impressively once. The capability that dazzles in a demonstration frequently collapses in production, and the collapse is not a sign of a bad vendor so much as a sign of an immature capability being sold as a mature one.

Why agents in particular struggle

Agentic systems are especially exposed to this gap, for a reason that is arithmetic as much as technical, illustrated in Figure 7. An agent that performs a multi-step task must complete every step correctly to succeed, and the reliability of the whole is the product of the reliability of each step. Even a step that is ninety-five percent reliable, which sounds impressive, compounds ruinously over a long sequence: across a dozen steps, the chance of a fully correct result falls to little more than half. This is why an agent can perform beautifully in a three-step demonstration and fail routinely in a twelve-step real workflow. Vendors mitigate the problem with checks, retries, and human approval points, and those techniques truly help, but they do not eliminate the underlying pressure, and they often reintroduce the human effort that the autonomy was supposed to remove. When a vendor shows an agent completing a task, the essential question is how many steps the real task requires and what the reliability is across all of them, not how smooth the demonstration looked.

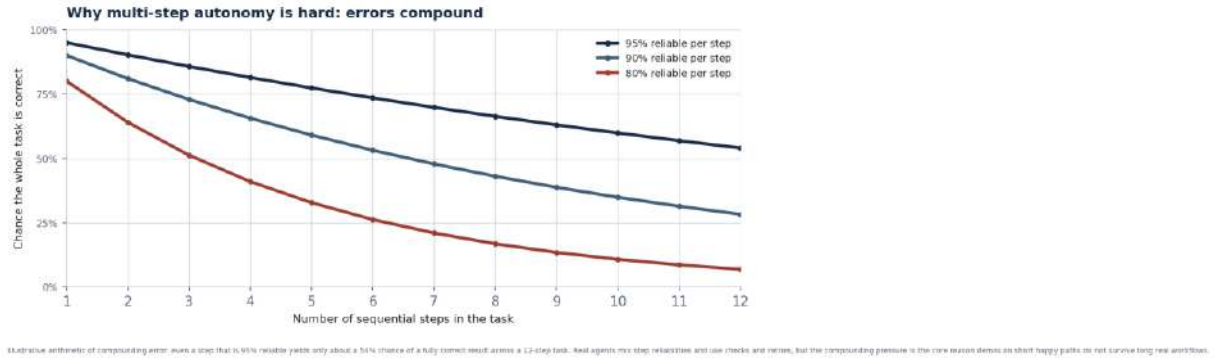

Figure 7. How per-step reliability compounds over a multi-step task. High per-step reliability still yields low end-to-end reliability across long sequences, which is why demos outrun production.

Beyond compounding error, agentic and generative systems carry a set of production challenges that demonstrations rarely reveal. They can hallucinate, producing confident output that is wrong, which is tolerable in a brainstorm and dangerous in an execution system. They are brittle at the edges, handling the common case well and the unusual case poorly, and supply chain is full of unusual cases. They struggle with the last mile of enterprise integration, connecting reliably to the specific, quirky systems a business actually runs. And they require human oversight to catch their errors, which means the promised autonomy is frequently a human-in-the-loop reality in which a person must review what the

system produces. None of these is disqualifying, and all are being worked on, but each is a reason the demonstration overstates the production capability, and a buyer who evaluates on the demonstration alone is evaluating the capability at its most flattering and least representative.

The last mile of enterprise integration deserves particular emphasis, because it is where supply chain agents most often falter and where demonstrations are least representative. A demonstration connects to a clean, cooperative version of the systems an agent must work with; production connects to the real ones, with their undocumented behaviors, their inconsistent data, their outages, and their edge cases accumulated over years. An agent that plans a sequence of actions must execute each against these real systems, and the messiness that a demonstration hides is exactly what breaks a long autonomous sequence in practice. This is why a capability can be real in the vendor's environment and fragile in the buyer's, and why the production references that matter most are those at the buyer's own scale and integration complexity, not the polished reference the vendor prefers to show.

Hallucination deserves particular caution in an execution setting, because its danger changes with the stakes. When a generative system invents a plausible but false answer in a brainstorm, the cost is trivial and the human catches it. When the same tendency appears in a system that acts, that reserves capacity, reallocates inventory, or commits an order, a confident error becomes a real transaction, and the compounding dynamic means one such error can propagate through subsequent steps before a human notices. This is why the move from an assistant that suggests to an agent that acts raises the reliability bar sharply, and why a capability that is acceptable as a suggestion engine may be unacceptable as an autonomous actor. A buyer evaluating an agent should ask specifically what the system does when it is uncertain, whether it can recognize the limits of its own confidence, and what stops a confident error from being executed.

09 The evidence that pilots stall

The gap between demonstration and production is not a theoretical concern; it is visible in the aggregate outcomes of enterprise AI programs, and the most comprehensive recent evidence is sobering. A widely discussed 2025 study from a research initiative at the Massachusetts Institute of Technology, examining hundreds of enterprise generative AI deployments alongside executive interviews and surveys, found that roughly ninety-five percent of enterprise generative AI pilots delivered no measurable impact on profit and loss, with only about five percent producing significant value. The authors named the pattern the divide between high adoption and low transformation: generic tools are adopted widely for individual productivity, while the enterprise-grade systems meant to change business outcomes are quietly abandoned. The drop is stark, as Figure 8 shows: of organizations that evaluated enterprise generative AI tools, a majority evaluated, a fifth reached a pilot, and only a twentieth reached production.

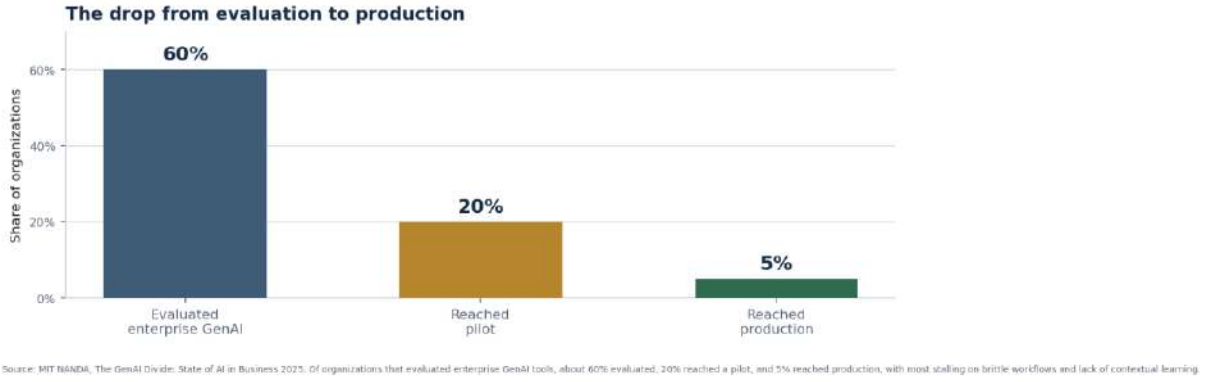

Figure 8. The attrition from evaluation to production in enterprise generative AI, as documented by MIT. Most initiatives stall well before they change any business outcome.

The study's diagnosis of why is directly relevant to AI washing, because it locates the failure not in the models but in the gap between what was sold and what was needed. The systems that failed did so, the researchers found, primarily because they were brittle in real workflows, could not retain feedback or context, and did not improve over time, which is to say they performed in the demonstration and collapsed in the operation. This is the production gap measured at scale. It is also a caution against the specific form of AI washing that presents a polished pilot as evidence of production readiness, since the data show that the journey from a pilot that impresses to a system that delivers is the journey most initiatives fail to complete.

What the survivors did differently

The same body of research offers a constructive counterpoint that should shape how buyers act, shown in Figure 9. The initiatives that succeeded shared a pattern, and two elements stand out. First, buying from specialized external vendors succeeded roughly twice as often as building internally, because the vendors brought domain-specific systems designed to fit real workflows rather than general tools that had to be adapted. Second, the successful buyers evaluated tools on business outcomes rather than on demonstrations or benchmarks, demanded deep customization to their own processes, and held

vendors accountable to measured results. The research also documented a striking shadow economy in which workers at the great majority of firms use personal AI tools even where official pilots have failed, a sign that the value is real where the fit is right and that the failures are failures of application, not of the underlying technology.

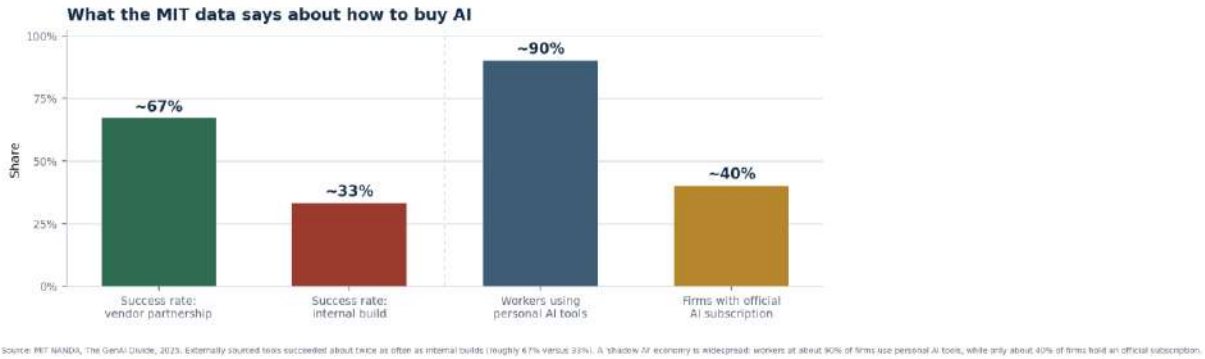

Figure 9. What distinguished success in the MIT data: external, workflow-fitted tools outperformed internal builds, and personal AI use far outran official adoption.

The lesson for a buyer evaluating AI claims is therefore double-edged and precise. The high failure rate is a warning against believing the demonstration and the marketed maturity, because most initiatives that begin with both still fail. But the success pattern is a guide to buying well: choose vendors whose systems are built for the specific workflow, insist on evaluation against business outcomes rather than demonstrations, and require the accountability that separates the five percent that deliver from the ninety-five that do not. Seeing through AI washing and buying AI successfully turn out to be the same discipline.

One articulation of why brittle AI fails in practice has become known as the verification tax. If a system is confidently wrong even a small fraction of the time, and does not signal when it is wrong, then a human must check every output to catch the errors, and the checking can consume more time than the system saved. The productivity promised by the demonstration evaporates into the labor of verification, and the capability that looked like a time-saver becomes a time-sink. This dynamic explains much of the gap between pilot enthusiasm and production disappointment: the tool works often enough to impress and unreliably enough to require constant supervision, and the supervision is where the promised savings go. A buyer should therefore measure not only how often a capability is right but whether it knows when it is wrong, because a system that flags its own uncertainty can be trusted to act where one that is confidently wrong cannot.

10 How vendors exaggerate: a field guide

AI washing follows recognizable patterns, and a buyer who knows the tactics can spot them in a pitch, a demonstration, or a written response. The following is a field guide to the most common forms, offered not to impugn every vendor, most of whom are describing real capability, but to arm the buyer to tell the real from the inflated.

1. **Agent washing and rebranding.** The relabeling of chatbots, robotic process automation, rules engines, and assistants as agentic AI, cognitive, or self-driving, without the underlying autonomy. The tell is that the capability described is one the buyer could have bought under a plainer name a year ago.
2. **The curated demonstration.** A demonstration that runs only on clean, prepared data and follows the intended path, showing the capability at its best and never at its edges. The tell is reluctance to run the demonstration on the buyer's own messy data or to show what happens when an input is unexpected.
3. **The vague claim.** Language such as powered by AI, AI-driven, or cognitive, with no specifics about what decision the AI makes, on what data, using what kind of model, at what level of autonomy. The tell is that the claim survives no follow-up question, dissolving into generality when pressed for detail.
4. **Roadmap presented as reality.** Capabilities that are announced, in limited preview, or available only to design partners, described in language that implies they are shipped and proven. The tell is the word soon, or its absence where it belongs, and the difficulty of finding a customer using the capability in production.
5. **Accuracy and return claims without method.** Impressive figures for accuracy, savings, or return on investment, presented without a baseline, a methodology, or independent validation. The tell is that the number has no denominator: better than what, measured how, verified by whom.
6. **Borrowed credibility.** A partnership with a well-known foundation-model provider presented as though it were proprietary capability, so that the buyer credits the vendor with the sophistication of the underlying model. The tell is that the vendor's actual contribution, as opposed to the model it calls, is never clearly described.

The common structure across all six is a claim that cannot survive a specific question. Vague language, curated demonstrations, unverified numbers, and borrowed credibility all collapse the moment a buyer asks precisely what the system does, on what data, with what autonomy, proven where. That is why the diligence questions in the next section are the single most effective defense against AI washing: not because vendors are dishonest, but because a specific question forces a specific answer, and a specific answer is exactly what an inflated claim cannot provide.

Setting the washed claim beside its honest equivalent makes the tactics concrete. Where a washed pitch says self-driving supply chain, an honest one says machine-learning forecasting with automated replenishment within set rules and human approval above a threshold. Where a washed pitch says cognitive, autonomous agents, an honest one says a conversational assistant that retrieves and drafts, with agent features in limited preview. Where a washed pitch offers ninety-nine percent accuracy, an honest one states an accuracy figure on a named data set, measured against a named baseline, validated by a named method. The washed version in each pair is not always false; it is unspecific, and

its unspecificity is the room in which exaggeration lives. The honest version answers the questions the washed version evades, and a buyer can often convert one into the other simply by asking, which is the fastest way to learn whether the specificity exists behind the slogan.

11 The questions that expose AI washing

The most powerful tool a buyer has is a list of specific questions asked consistently of every vendor, because AI washing thrives on generality and dissolves under specificity. The questions below are organized by what they test, and the value of each is that a real capability produces a clear, confident, evidenced answer while an inflated one produces hedging, redirection, or a return to marketing language. A buyer need not be a data scientist to use them; the quality of the answer, its precision and its evidence, is itself the signal.

What the system actually is

- What exactly is the AI doing, which specific decision does it make or support, and on what data? A real answer names a decision and a data set; an inflated one describes a vibe.
- Which rung is it on: fixed rules, classical machine learning, a generative model, or an autonomous agent? Vagueness here is the central tell of AI washing.
- Is the model yours, a third party's, or a combination, and what specifically does your software add beyond the underlying model?

What autonomy it actually has

- Does the system recommend, or does it act? If it acts, within what bounds, and what does it do without a human approving the step?
- What is the human-in-the-loop design, and how much human effort does the capability actually require in practice, as opposed to in the demonstration?
- For a multi-step task, what is the measured reliability across all the steps, and what happens when a step fails?

What is proven, and where

- Is this in production with real customers at my scale and in my industry, or is it a pilot, a preview, or a roadmap item? Ask for the distinction explicitly.
- What is the measured accuracy or performance, against what baseline, and validated by whom? A number without a baseline is not evidence.
- Can I see it run on my own data rather than the demonstration data, and can I speak to reference customers running it in production?

How it is governed

- How is the system monitored, how are its errors detected and corrected, and how does it improve over time?
- Where does my data go, is it used to train models that benefit competitors, and what are my data rights and portability terms?

- What happens on the edge cases and the failures, and what is the fallback when the AI is wrong or uncertain?

Reading the written response

The same scrutiny applies to written responses to a request for proposals, where AI washing hides in prose as readily as in a demonstration. The tells in writing are specific. Watch for capability described in adjectives rather than mechanisms, intelligent, cognitive, autonomous, with no account of what decision is made on what data. Watch for the passive voice that hides the actor, decisions are optimized rather than the system performs this specific action. Watch for numbers without denominators, and for the present tense applied to capabilities that a careful reading reveals to be planned. And watch for answers that describe the underlying foundation model's abilities rather than the vendor's own, which borrow credibility the vendor has not earned. A useful discipline is to require, for every AI claim in a response, a single sentence naming the decision, the data, the autonomy level, and a production customer, because that sentence is easy to write about a real capability and nearly impossible to write about an inflated one.

> **The single most revealing request**
>
> - **Show me on my data.** Ask the vendor to run the capability on your own messy data, at your scale, including your hard cases, rather than on the demonstration set. A real, mature capability welcomes this; an inflated one finds reasons to defer it. No single request separates the real from the washed more reliably.

Using these questions well is a matter of discipline more than expertise. Ask the same questions of every vendor, record the answers, and compare them side by side, because the contrast between a precise, evidenced answer and a vague, redirecting one is far more revealing than either answer alone. Bring someone who understands the technology to the room if possible, but do not defer the judgment to them entirely, because the signal a buyer is reading is not the technical detail itself but whether the vendor can supply it clearly and back it with evidence. And return to the unanswered questions rather than letting them slide, because the questions a vendor cannot answer cleanly are precisely the ones whose answers the buyer most needs. The questions are simple; the value is in asking them consistently and in refusing to accept marketing language where a specific answer belongs.

12 Designing a proof of concept that tests reality

Questions expose inflated claims; a well-designed proof of concept proves real ones. The purpose of a proof of concept is not to watch the vendor demonstrate the capability, which proves only that a demonstration exists, but to test the capability under conditions that resemble production, which is the only way to know whether it will survive there. A proof of concept designed to flatter the vendor is worse than none, because it manufactures false confidence; a proof of concept designed to stress the capability is the most valuable diligence a buyer can perform.

The principles of an honest proof of concept

- **Run it on your data, not theirs.** Use your own real, messy, representative data, including the difficult cases, rather than the vendor's curated set. Capability that works only on prepared data is not capability you can use.
- **Measure against a baseline.** Define, before the test, what the current process achieves, and measure the AI against it. An improvement claim is meaningless without the baseline it improves upon, and many capabilities that impress in isolation add little over what the organization already does.
- **Test the edges and the failures.** Deliberately include the unusual, the malformed, and the adversarial, because the common case is where every system performs and the edge case is where the real difference lies. Observe not just whether it fails but how it fails, and whether the failure is safe.
- **Measure the human effort actually required.** Track how much human review, correction, and intervention the capability truly needs in operation, because the gap between the promised autonomy and the required oversight is where the value is won or lost.
- **Run it long enough to see reliability, not just possibility.** A capability that works once is not the same as one that works reliably across the volume and variety of real operation. Run the proof of concept long enough, and across enough cases, to observe consistency rather than a single success.

A proof of concept built on these principles does something a demonstration never can: it converts a claim into evidence specific to the buyer's own environment. It also has a useful secondary effect, which is that vendors confident in a real, mature capability engage with such a test readily, while vendors whose capability is inflated resist it, negotiate its terms, or steer it back toward the curated demonstration. The willingness to be tested openly is itself a strong signal, and the design of the test is where a buyer converts the skepticism this guide counsels into a decision grounded in fact.

A concrete example shows the design at work. Suppose a buyer is evaluating a vendor's claim of superior exception detection for inbound shipments. A flattering demonstration would show the system flagging obvious problems on the vendor's prepared data. An honest proof of concept would instead run the system on a defined historical period of the buyer's own shipment data, for which the true exceptions are already known, and measure how many real exceptions it caught, how many false alarms it raised, and how its performance compared to the buyer's current process over the same period, including on the unusual cases that matter most. That test converts the claim into a number the buyer

can trust, exposes whether the capability adds value over what the organization already does, and reveals the false-alarm burden that determines whether the capability is usable in practice. It is more work than watching a demonstration, and it is the difference between buying a capability and buying a claim.

13 Why AI washing is a procurement and risk problem

It would be easy to treat AI washing as a marketing annoyance to be shrugged off, but that underestimates it. Inflated AI claims impose real costs and real risks on the organizations that fail to see through them, and those consequences make AI washing a procurement and governance problem that deserves board-level attention, not merely a buyer's irritation.

The most direct cost is financial. A buyer who believes a conventional system is an advanced AI pays an AI premium for ordinary software, and a buyer who stakes a program on a capability that is not yet real funds an implementation that cannot deliver what was promised. The failure statistics examined earlier are, in part, a measure of this cost aggregated across the economy: the enormous sums invested in pilots that never reach production represent, in many cases, capabilities that were sold as more mature than they were. Beyond the direct waste lies opportunity cost, the value of the better-fitted or more honest solution that was not chosen because an inflated one won the evaluation, and the time lost to a program that must be restarted when the capability proves inadequate.

The risks extend past cost into compliance and reputation. For public companies, the regulatory turn described earlier means that repeating a vendor's inflated AI claims in the organization's own disclosures can carry the same liability the vendor faces, a risk that has moved from theoretical to enforced. Operationally, deploying an AI capability that is less reliable than believed, particularly an agentic one that acts rather than recommends, can cause real damage, from mis-executed transactions to eroded customer trust, and the compounding-error dynamic examined earlier means that the failures of an over-trusted agent can cascade before anyone notices. And there is a governance risk in the data terms that often accompany AI features, under which a buyer's proprietary data may be used to train models that benefit competitors, a cost that is invisible at signing and significant over time.

Framed this way, seeing through AI washing is not skepticism for its own sake but ordinary procurement diligence applied to an area where the information asymmetry is unusually large and the hype unusually loud. The buyer who treats AI claims with the same rigor applied to any other material representation, demanding specifics, evidence, and accountability, protects the organization from paying for capability it will not receive and from staking its operations on capability that does not yet exist. That rigor is best embedded not in the judgment of a single evaluator but in the procurement process itself, which is the subject of the final section.

A concrete cost illustration shows why the stakes justify the diligence. Suppose an organization selects, on the strength of an impressive agentic demonstration, a system priced at a premium over a conventional alternative, and commits to a substantial implementation on the expectation of autonomous operation. If the autonomy proves to be a human-in-the-loop reality that requires the oversight the demonstration hid, the organization pays three times: the premium over the conventional system it could have bought, the implementation cost of a program that does not deliver the promised savings, and the opportunity cost of the year lost before the shortfall is clear and the effort must be restarted. None of these costs appears in the evaluation that chose the system, and all of them follow

from believing a claim that a proof of concept on the organization's own data would have tested. The diligence this guide prescribes is inexpensive against the cost of skipping it.

The disclosure dimension deserves a further word, because it has become a live risk rather than a hypothetical one. Public companies increasingly describe their use of artificial intelligence in investor communications, and a company that repeats a vendor's inflated claim, describing as autonomous or intelligent a capability that is neither, may find that it has made a misleading statement of its own. The regulatory actions described earlier were brought against companies for their own AI claims, and the same standard applies to any organization that describes its capabilities to investors or customers. A buyer who has done the diligence this guide prescribes, and who therefore knows what its systems actually do, is also protected on the disclosure side, because it can describe its capabilities accurately. The diligence that prevents overpaying for capability also prevents overstating it, which is a second and increasingly valuable return on the same effort.

14 Requirements, acceptance criteria, and a scoring rubric

The defenses in this guide are most powerful when they are built into how an organization buys, rather than left to the discernment of whoever happens to run an evaluation. Three practices embed the discipline in procurement itself.

Define the capability, not the label

Write requirements that specify what the system must actually do, in terms of decisions, data, autonomy, and measured performance, rather than requirements that ask for AI or agentic capability in the abstract. A requirement that asks for an agentic planning assistant invites AI washing; a requirement that asks for a system that detects a specified exception, on specified data, at a specified accuracy against a specified baseline, with a specified level of autonomy, cannot be satisfied by a relabeled chatbot. The act of writing the requirement in concrete terms is itself a filter, because it forces the buyer to decide what capability is truly needed and gives the evaluation an objective standard.

Require production evidence and tie payment to performance

Make production evidence a condition of the purchase, not a nicety: require reference customers running the capability at comparable scale, and make a successful proof of concept on the buyer's own data a gate that must be passed before commitment. Then tie payment and acceptance to measured performance, so that the vendor is paid for the capability delivered rather than the capability described. Acceptance criteria written around measured results, the accuracy achieved on the buyer's data, the reliability across the real task, the human effort actually required, convert the vendor's claims into contractual obligations, and a vendor confident in a real capability will accept such terms while one selling an inflated capability will resist them.

A scoring rubric

For teams comparing vendors, the dimensions below can be scored consistently, with the pattern rather than any single score guiding the decision. The rubric operationalizes the guide, forcing each vendor to be assessed on substance rather than on the impression left by a demonstration.

Dimension	What a strong answer looks like	Red flag
Capability clarity	Names the decision, data, and model type	Vague 'AI-driven' language
Level on the ladder	Honest about rules, ML, GenAI, or agent	'Agentic' with no autonomy shown
Autonomy reality	Clear on what it does without a human	Recommends but is sold as acting
Production evidence	Reference customers at your scale	Only pilots, previews, or roadmap
Measured performance	Accuracy against a baseline, validated	Numbers with no method or baseline
Proof on your data	Welcomes a test on your messy data	Defers or resists your-data testing
Data rights and governance	Clear terms; your data stays yours	Trains on your data for others

Read the rubric as a pattern. A vendor that answers clearly on capability and autonomy, sits candidly on the ladder, brings production evidence and measured performance, welcomes a test on the buyer's own data, and offers clean data terms is a vendor selling something real. A vendor that trips the red flags across several dimensions is selling a claim, however polished the demonstration, and the rubric earns its keep when it stops an inflated purchase that a compelling pitch would otherwise have won.

The evidence on how successful buyers behave also speaks to a question many organizations face directly: whether to build AI capability internally or buy it from a specialized vendor. The research examined earlier found that externally sourced tools succeeded roughly twice as often as internal builds, not because buying is inherently superior but because specialized vendors bring systems already fitted to real workflows, while internal builds routinely underestimate the cost of integration and the difficulty of the last mile. For most supply chain organizations, whose core competence is running a supply chain rather than building AI, the implication is to buy from vendors with real, proven capability and to concentrate internal effort on the integration, the data, and the workflow fit that determine whether a bought capability delivers. The discipline is the same either way: define the capability, demand the evidence, and measure the result, whether the capability is built or bought.

15 Conclusion: reward the real

The argument of this guide is easy to mistake for cynicism about artificial intelligence, and it is the opposite. AI is one of the most consequential technologies to reach the supply chain in a generation, and in demand forecasting, anomaly detection, optimization, vision, and document processing it delivers substantial, measurable value today. The move toward agentic systems that can act within bounds is a real advance with real promise. The problem this guide addresses is not the technology but the marketing, which has run so far ahead of the reality that the label AI has lost its power to tell a buyer anything, and that a large share of what is sold as intelligent or autonomous is neither. In that environment, the buyer's task is not to reject AI but to discriminate: to find and reward the real, and to decline to pay for the inflated.

Doing that requires holding two truths at once, which is the honest posture the evidence supports. AI is real and valuable, and most of what is marketed as AI is more conventional, less autonomous, and less proven than its packaging implies. Regulators have begun to enforce against the gap, analysts have named it, and the majority of pilots built across it fail to reach production. A buyer who believes only the first truth overpays and is disappointed; a buyer who believes only the second misses the genuine value and falls behind. The discipline this guide describes, understanding the ladder from rules to agents, mapping where AI is and is not real, recognizing the tactics of exaggeration, asking the specific questions that inflated claims cannot answer, and testing capability on one's own data before committing, is the way to hold both truths and act on them.

The deepest point is that seeing through AI washing and buying AI well turn out to be the same skill. The research on why pilots fail and why some succeed points to the same practices the diligence in this guide prescribes: choose capability built for the real workflow, insist on evidence over demonstration, measure against a baseline, and hold the vendor accountable to results. The buyers who master this will not be the ones who avoid AI, nor the ones who believe every claim, but the ones who can tell the difference, and who structure their evaluations and contracts to reward the vendors building something real. In a market crowded with the marketed, the ability to find the real is a durable advantage, and it belongs to the buyer who asks, of every impressive claim, the one question that inflated capability can never answer well: show me, on my data, that it works.

16 Methodology, caveats, and sources

Methodology

- This article synthesizes regulatory actions, analyst research, an academic study of enterprise AI adoption, vendor disclosures, and technology reporting, current to mid-2026. Supply Chain Research is independent and accepts no payment from the vendors or firms discussed.
- Where vendor capabilities are described, they are drawn from public announcements and should be verified directly, because this space moves quickly and the distinction between shipped, piloted, and announced capability is exactly what a buyer must confirm.

Caveats

- This is a fast-moving field, and specific figures and vendor capabilities will change. The agentic AI landscape in particular is early, and a capability that is a laboratory demonstration at the time of writing may be in production shortly after, or may not. Every claim, including the vendors' and the analysts', should be checked against current evidence.
- The headline statistics come from different sources measuring different populations and should not be combined. The MMC Ventures figure on European AI startups is from 2019 and measured a specific sample; the MIT figure on pilot failure and the Gartner figures on agentic vendors and cancellations are from 2025 and use their own definitions and methods. They are corroborating indicators of a pattern, not a single measurement.
- Several figures are analyst predictions or survey-based estimates rather than audited facts. The forty percent cancellation figure is a Gartner forecast; the count of roughly one hundred thirty real agentic vendors is a Gartner estimate; the trust and adoption figures are survey results. They are directional and are presented as such.
- The maturity map of supply chain AI use cases is a directional assessment, not a benchmark, and placement varies by vendor and specific application. The compounding-error illustration is simplified arithmetic meant to convey a dynamic, not a measurement of any particular system.

Sources

[1] U.S. Securities and Exchange Commission (March 18, 2024). SEC charges two investment advisers for making false and misleading statements about AI.

[2] U.S. Federal Trade Commission (September 25, 2024). FTC announces crackdown on deceptive AI claims and schemes (Operation AI Comply).

[3] Gartner (June 25, 2025). Gartner predicts over 40% of agentic AI projects will be canceled by end of 2027.

[4] MIT NANDA (2025). The GenAI Divide: State of AI in Business 2025.

[5] U.S. Federal Trade Commission. DoNotPay, cases and proceedings.

[6] Kinaxis (October 2025). Kinaxis accelerates the agentic era for supply chain orchestration with Maestro Agents.

[7] SiliconANGLE (June 2026). Supply chain operations turns agentic: Blue Yonder on agents in production and in the lab.

[8] Gartner. Agentic AI and the practice of 'agent washing' (newsroom and analyst commentary).

Additional context drawn from the widely cited MMC Ventures State of AI report (2019) on the share of European 'AI' startups with no material evidence of AI; from Workato and Harvard Business Review research on the gap between agentic AI investment intent and trust in autonomy; and from vendor disclosures by planning, warehouse, transportation, and visibility providers. Vendor capabilities should be verified directly, as

the distinction between shipped, piloted, and announced capability changes rapidly. Analyst predictions and survey figures are directional. This article is analysis, not legal, procurement, or investment advice.
