Agentic AI Governance: Before an AI Agent Gets Authority to Act

An Executive Control Framework for Authorizing, Limiting, and Auditing AI Agents

David W. Koran

CyberAB Registered Practitioner Advanced

ISO/IEC 27001 Auditor and Implementer

August 2026

Summary

The question that separates one generation of business AI from the next is not whether a company uses artificial intelligence but whether an AI system holds authority to act. A tool that drafts text for a person to review and send operates under a different set of concerns than a tool that sends the message itself, schedules the delivery, executes the payment, or changes the record. The latter kind of system is an agent, and an agent requires a form of governance that most organizations have never applied to software: the deliberate grant, limitation, and audit of authority.

This paper gives management a control framework to apply before that authority is granted. The framework consists of eight decisions: what the agent may read, what actions it may take, where human approval is required, what identity and privileges it operates under, how its actions are logged and reconstructed, what the vendor is responsible for, how the agent is shut down and its actions reversed, and how its authority expires and is renewed. Each decision is an extension of controls that well-run organizations already apply to people: access authorization, signature authority, segregation of duties, and accountability for actions taken. The decisions themselves draw on familiar management disciplines, although legal, cybersecurity, and technical specialists may be needed to verify some of the answers.

The urgency comes from how agents arrive. For many small and midsize organizations, the first AI agent will not come through a procurement decision that triggers a review. It is at least as likely to arrive inside a tool the company already approved, delivered by a vendor update that may enable the capability by default. A company that has made the eight decisions in advance will handle that moment with a meeting and a concise authority record, while a company that has not will discover its first agent only after the agent has acted.

What Makes an Agent Different

The most visible recent wave of business AI has been generative. A person asks for a draft, a summary, an analysis, or an image, and the system produces content. Whatever the quality of that content, a person stands between the output and the world. The person decides whether the draft is sent, whether the analysis is relied upon, whether the summary enters the record. Governance of generative AI is therefore mostly governance of data and judgment: what information goes into the tool, and how much weight its output is given.

An agent can partially or completely remove the person from that position for specified actions. An agentic system reads information from connected systems, makes a decision within a defined scope, and executes an action with real-world effect: it sends the message, books the appointment, places the order, updates the record, or initiates the payment. For purposes of this paper, an AI agent is an AI-enabled system that can interpret a goal or trigger, select or sequence actions using connected data and tools, and carry out some of those actions with limited human supervision. Traditional automation can also act, but it generally follows predetermined rules. The property that matters for governance is

authority rather than intelligence, and the trigger for the framework in this paper is delegated authority: the point at which software has been given, deliberately or by default, the ability to alter the state of the business without a person touching the change.

That shift changes the governance question. For generative tools, the question is whether the output is reliable and where the data goes. This firm's earlier paper on deploying AI in a defense manufacturing business organized that analysis by data class, staging adoption according to where the data travels (davidkoran.com/ai-governance/white-papers/ai-in-defense-manufacturing/). Agent governance adds a second axis. Alongside the question of what the system may see stands the question of what the system may do. A capable organization answers both, and answers them before the system operates.

The closest existing analogy in most companies is signature authority. No well-run business lets a new employee sign contracts, issue payments, or commit delivery dates on their first day. Authority is granted in defined amounts, for defined purposes, with defined approvals above a threshold, and the grant is recorded. An AI agent deserves exactly the same treatment, with one addition: unlike an employee, an agent will exercise its authority at machine speed and volume, so a badly scoped grant compounds faster than any human error could.

How Agents Arrive

Agents reach a company by three paths, and the path determines whether governance happens at all.

The first path is deliberate procurement. Management identifies a process, evaluates agentic products, negotiates terms, and deploys. This is the path every governance framework quietly assumes, because it contains a natural decision point, and it is the path that arrives with the clearest checkpoint, which is precisely what the other two paths lack.

The second path is the vendor update. A tool the company already uses, already approved, and already trusts ships a new release, and the release includes an agentic capability: an assistant that can send on a user's behalf, a scheduling feature that can commit the calendar, an accounting integration that can execute a payment, a customer service function that can answer and resolve without review. The capability may arrive enabled by default or without any new approval step, announced in release notes few people read, and presented as a convenience rather than a transfer of authority. For small and midsize organizations this path deserves the most attention, because it bypasses every existing checkpoint: the tool was approved before it could act, and no procurement review will ever be triggered because nothing was procured.

The third path is unapproved adoption. An employee connects a personal automation service or browser-based agent to company accounts to save time. This path is a species of shadow AI, and the organizational answer is the same one that governs unapproved tools generally: a clear use policy, an approved alternative, and monitoring proportionate to the environment. The remainder of this paper concentrates on the first two paths, where the company itself is making, or failing to make, the decision.

The practical conclusion from the three paths is a standing rule rather than a one-time project: no system, however it arrives, exercises authority to act until the decisions in the next section have been made and recorded. The rule can be adopted without purchasing anything, and it converts the vendor update path from a surprise into a routine review.

The Control Framework

The framework below is deliberately built from management disciplines the reader already operates. Access authorization, spending limits, approval thresholds, segregation of duties, audit trails, and periodic review are not new inventions for AI. They are how organizations already govern people and money. The work of agent governance is to apply them, explicitly and in writing, to a new category of actor. The eight decisions are consistent with risk management and governance themes in the NIST AI Risk Management Framework and ISO/IEC 42001, both of which are treated at length elsewhere on this site, but they are not a substitute for full framework implementation or certification where an organization pursues either. They are also consistent in direction with the agent-specific standards work NIST opened in 2026, which the sections on identity and logging below take up.

1. Read Authorization

The first decision defines what the agent may see: which data, in which systems, at what level of sensitivity. The scope should be stated positively, as a list of what is authorized, rather than negatively, as a list of exclusions, because vendor connectors expand over time and a positive list fails safely. Data class does the heavy lifting here. An agent authorized to read public product information and open order status is a different risk than an agent authorized to read contract files, personnel records, or regulated technical data, and the authorization should say so in those terms. For defense contractors, the rule from staged deployment carries over with one refinement, because the two categories of regulated data are not governed identically. No agent reads Controlled Unclassified Information (CUI) until the agent, its connected systems, any service providers involved in its operation, and its data paths have been scoped under DFARS 252.204-7012 and the CMMC scoping requirements at 32 CFR 170.19. Export controlled technical data requires a separate analysis under the applicable export control regime, because access and disclosure questions under the ITAR (22 CFR 120.33) and the EAR (15 CFR 734.15) turn on who can receive the information, not only on which system holds it.

Read authorization carries a second risk that generative tools do not present in the same degree. Information the agent reads from outside the organization must be treated as untrusted input, because emails, attachments, web pages, and customer documents may contain instructions designed to redirect the agent or cause it to misuse an authorized capability, a technique known as prompt injection. Authorizing an agent to read a source is therefore also a decision about action risk, and the review must consider whether content from that source can influence what the agent does, not only what the agent learns.

2. Action Authorization

The second decision defines what the agent may do, and it is the decision most organizations skip because vendors present capabilities as features rather than authorities. The authorization should enumerate the transactions and communications the agent may execute, and it should state financial and operational limits as numbers, not adjectives. An authorization that permits the agent to send routine confirmations leaves every meaningful boundary undefined, while an authorization that permits autonomous order status confirmations on orders under $5,000, capped at 25 messages per day, gives every downstream control something to enforce. Numbers create the possibility of an exception, and exceptions are what logging, approval, and incident processes act on.

The action authorization is also the natural home for predeployment testing. Before the agent operates, the review should exercise it against representative business scenarios and against the failure cases that matter: attempts to exceed the financial and volume limits, content containing embedded instructions, requests for data outside the read scope, unavailable downstream systems, duplicate transactions, and attempts to route around the approval threshold. An authorization granted without testing is a limit on paper rather than a limit in operation.

3. Approval Thresholds

Between full autonomy and full review lies the threshold: the defined line above which an agent′s proposed action waits for a named human approver. Setting the threshold is a management judgment about consequence, not a technical judgment about model quality. A useful test is reversibility. Actions that are cheap to reverse can sit below the threshold; actions that commit the company, in money, in schedule, or in words a customer will rely on, belong above it. The threshold holder must be named, must have time to actually review, and must not be so burdened that approval becomes a reflex, because an approval given without reading provides none of the review the threshold was created to guarantee.

4. Identity and Privilege

An agent must act under a distinct and attributable identity that carries the least privilege necessary for the authorized scope. Depending on the architecture, that identity may be implemented as a dedicated service account, a workload identity, or a short-lived delegated credential. What is never acceptable is a shared credential or an inherited user session, which destroys accountability twice over: the log cannot distinguish the agent′s actions from the person′s, and the agent silently acquires every privilege the person holds. The authorization should also name the business owner responsible for the agent and, where the agent acts on a person′s behalf, the person whose authority it exercises. These are the same questions NIST opened for public comment in early 2026 in its concept paper on software and AI agent identity and authorization, which addresses agent identity, delegated authority, and the binding of human approvers to agent actions. Segregation of duties applies with full force. An agent that initiates a transaction must not

be the mechanism that approves it, and an agent must never hold the privilege to modify the logs that record its own behavior.

5. Logging and Reconstruction

The standard for agent logging is reconstruction: after the fact, management must be able to establish the triggering event, the governing instruction, the information sources consulted, the tool calls made, the authorization decision, the human approval where one applied, the model and version, the action taken, the outcome, and any exception generated. Each action should be attributable to the agent's identity and, where a human released it, to the releasing user, and the logs must be exportable, retained for an appropriate period, and protected from alteration by the agent itself. The control should not depend on access to hidden model reasoning. It operates on the observable evidence trail at the system boundary, which is also where NIST's early agentic evaluation research concentrates, building structured audit trails that connect an agent's outputs to the evidence behind them. A capability that cannot produce that account at the boundary fails the review, whatever else it offers.

Logging is retrospective by nature, and a monitoring layer is what makes the evidence operational. The review should define the conditions that generate an alert rather than merely a record: actions outside authorized scope, abnormal transaction volume, repeated rejections at the approval threshold, the appearance of a new tool or connector, and any change in the model or the vendor configuration. NIST's March 2026 report on the challenges of monitoring deployed AI systems treats post-deployment monitoring as a discipline in its own right, and its direction supports the same conclusion: evaluation of an operating system does not end at approval.

6. Vendor Responsibility

Most agents arrive inside vendor products, which makes the contract a control. Before authority is granted, management should establish in writing what the vendor warrants about the agent's behavior, whether company data is used to train models, what the vendor commits to when the agent malfunctions, how capability changes are announced, and where responsibility divides between the vendor's operation of the model and the company's configuration of its authority. A vendor that cannot answer these questions in an order form, an addendum, or published terms is asking the company to grant authority to a system nobody stands behind.

7. Shutdown and Rollback

Every grant of authority needs a tested way to withdraw it. Before an agent operates, management should know the specific mechanism that stops it immediately, whether that is a feature toggle, a credential revocation, or a connector removal, and should have tested that the mechanism works. Rollback deserves the same forethought: which of the agent's actions can be reversed, by what procedure, and which cannot. A sent message cannot be unsent, and a payment may be recoverable only through the bank. Knowing the irreversible actions in advance is itself a scoping input, because irreversibility is a strong

argument for placing an action above the approval threshold. The time to design and test shutdown is before deployment, because an incident is the worst possible moment to discover that the procedure does not work.

8. Reauthorization and Incident Reporting

Authority should carry an expiration date, because a standing grant that is never reviewed drifts away from the conditions under which it was made: vendors add capabilities, connected systems change, and the business changes around the agent. The cycle should be proportionate to risk. Six months is a practical starting point for an agent holding meaningful authority, with immediate review following any material change to the model, the tools, the data sources, the permissions, or the vendor terms, and the cycle creates a natural moment to revisit declined capabilities, thresholds, and limits. Incidents follow the same principle of using what exists: an agent action outside its authorized scope is an incident or control exception that must be classified and routed to the applicable process, which depending on the action may be cybersecurity, privacy, legal, financial, quality, or customer response, and where the environment is regulated it is evaluated against the same reporting obligations as any other incident. Organizations with a corrective action process should route agent errors there as well, because the risk of a wrong decision recurring, in the same or a related form, remains until the configuration, the scope, the data, or the tool changes.

The Worked Example

Consider the same 50-person precision machining business that has appeared throughout this series: a defense subcontractor with commercial customers, an approved quoting and order management platform, a written AI use policy, and a staged deployment plan organized by data class. In August 2026, the platform vendor ships release 2026.3. The release notes announce an assistant that can draft and send customer communications on behalf of account managers, and an auto-reply capability for inbound customer email. Both are enabled by default.

Under the company's standing rule, the features do not operate until the eight decisions are made. The operations manager schedules a review, and the review takes one meeting. The team confines the agent's read scope to the platform's own customer, quote, and order records, keeping it away from email archives, file shares, accounting, and any repository connected to the CUI environment. It authorizes autonomous sending only for order status confirmations on orders under $5,000, capped at 25 messages per day, and places anything containing a price, a date commitment, or a contract term above the approval threshold, to be released by the account manager of record. It confirms the vendor provisions the assistant with a dedicated service account, verifies that the platform log captures message content, trigger, recipient, timestamp, model version, and releasing user, and tests the log export. It reviews the vendor's AI feature addendum, confirms customer data is not used for training, and files the addendum with the contract. It tests the feature toggle and the queue purge, and it runs the assistant through test scenarios, confirming that the volume cap holds, that a request for data outside the platform is denied, and that a message drafted from content containing embedded instructions is

held for review rather than sent. It declines the auto-reply capability entirely, documents the refusal, and sets February 2027 as the reauthorization date.

The output of the meeting is the record below, signed by the president and the operations manager and filed with the company's other authorization records.

AI Agent Authority Decision Record	
System and feature	Customer communication module of the approved quoting and order management platform; vendor release 2026.3 adds a send-on-behalf assistant
Date of review	August 12, 2026
Arrival path	Vendor feature update to a previously approved tool; no procurement action
Risk classification	Moderate: external customer communication with bounded financial exposure and a reversible sender queue
Owners	Business owner: Operations Manager; technical owner: IT coordinator, supported by the managed service provider
Read authorization	Customer contact records, open quotes, and order status in the platform only; no access to email archives, file shares, accounting, or any repository holding CUI or export controlled technical data
Action authorization	May draft outbound customer messages and place them in the sender queue; may send routine order status confirmations autonomously; may not issue quotes, accept orders, commit delivery dates, or communicate pricing
Limits	Autonomous sending limited to order status confirmations valued under $5,000 per order; maximum 25 autonomous messages per business day; all other drafts held for human release
Approval threshold	Any message containing a price, a date commitment, or a contract term requires release by the account manager of record; threshold reviewed at reauthorization
Identity and privilege	Agent operates under a dedicated service account with read scope above; no shared credentials; account cannot approve its own queued messages

AI Agent Authority Decision Record	
Logging	Vendor log verified to capture message content, triggering record, recipient, timestamp, model version, and releasing user; export tested on August 12, 2026
Monitoring	Alerts configured for actions outside authorized scope, daily volume above the cap, repeated messages held at the threshold, and vendor configuration changes
Testing completed	August 12, 2026: volume and value caps enforced; out-of-scope data request denied; message drafted from content containing embedded instructions held for review; approval threshold could not be bypassed
Vendor responsibility	Order form and AI feature addendum reviewed; vendor attests customer data is not used for model training; support commitment covers agent malfunction; addendum filed with the contract
Declined capability	Auto-reply to inbound customer email, enabled by default in release 2026.3, disabled and documented; to be reconsidered at reauthorization
Shutdown and rollback	Feature toggle disables the assistant immediately; sender queue purge reverses unsent items; sent messages handled through customer correction notice; procedure tested August 12, 2026
Reauthorization	February 12, 2027, or upon any material change to the model, tools, data sources, permissions, or vendor terms, whichever comes first
Incident reporting	Agent actions outside authorized scope classified and routed through the existing incident response and corrective action procedures
Approved by	President (signature on file); Operations Manager (signature on file)

Two features of the example deserve notice. First, nothing in the review required expertise in model design. The facts behind some answers came from the platform vendor and the managed service provider, but every decision the team made is a management decision about authority, and every control it applied already existed in the business in some form. Second, the review produced a refusal. The auto-reply capability failed not because the technology is poor but because the company saw no way to keep inbound email, which arrives with attachments, unpredictable content, and the possibility of instructions aimed

at the assistant itself, inside the read scope it was prepared to authorize. A governance process that produces no refusals over time is worth examining, since the pattern suggests the review is confirming decisions rather than making them. The declined capability waits, documented, for the reauthorization review, where it can be reconsidered against whatever the vendor has improved.

Before the First Action

The economics of agent governance depend entirely on timing. A framework applied before authority is granted costs a meeting and a concise authority record, while the same framework applied afterward costs an investigation: reconstructing what an unscoped agent did from whatever logs happen to exist, unwinding commitments the company never intended to make, and explaining to customers, insurers, or contracting officers why a system nobody reviewed was acting in the company′s name.

The organizations that handle the agentic transition well will not be the ones with the most sophisticated AI programs. They will be the ones that recognized authority as the thing being granted and treated it accordingly, with the same seriousness they already bring to signature authority, spending limits, and system access. That recognition is available to any organization at any size, and the only requirement is that it come before the first action, because once the agent has acted, the work changes from establishing a plan to conducting a cleanup.

About the Author

David W. Koran is a CyberAB Registered Practitioner Advanced (RPA) and the founder of a consulting practice serving Defense Industrial Base contractors and their legal counsel. His work focuses on CMMC readiness, enablement, and implementation. He holds ISO/IEC 27001 certifications as an auditor and implementer, is an Associate Member of the American Bar Association Section of Public Contract Law, and is the author of The CMMC Decision. He can be reached at dkoran@davidkoran.com or (802) 335-2662.

References

National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1, January 2023. AI RMF 1.0 remains the published framework as of August 2026 and is undergoing revision. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

National Institute of Standards and Technology, Center for AI Standards and Innovation, AI Agent Standards Initiative, launched February 17, 2026. https://www.nist.gov/artificial-intelligence/ai-agent-standards-initiative

National Institute of Standards and Technology, National Cybersecurity Center of Excellence, Accelerating the Adoption of Software and AI Agent Identity and Authorization, draft concept paper, February 2026. https://www.nccoe.nist.gov/sites/default/files/2026-02/accelerating-the-adoption-of-software-and-ai-agent-identity-and-authorization-concept-paper.pdf

National Institute of Standards and Technology, Building Evaluation Probes into Agentic AI. https://www.nist.gov/programs-projects/building-evaluation-probes-agentic-ai

National Institute of Standards and Technology, Center for AI Standards and Innovation, Challenges to the Monitoring of Deployed AI Systems, NIST AI 800-4, March 2026. https://doi.org/10.6028/NIST.AI.800-4

DFARS 252.204-7012, Safeguarding Covered Defense Information and Cyber Incident Reporting. https://www.acquisition.gov/dfars/252.204-7012-safeguarding-covered-defense-information-and-cyber-incident-reporting.

32 CFR 170.19, CMMC Scoping. https://www.ecfr.gov/current/title-32/subtitle-A/chapter-I/subchapter-G/part-170/subpart-D/section-170.19

22 CFR 120.33, Technical data (ITAR). https://www.ecfr.gov/current/title-22/chapter-I/subchapter-M/part-120/subpart-C/section-120.33

15 CFR 734.15, Release (EAR).
https://www.ecfr.gov/current/title-15/subtitle-B/chapter-VII/subchapter-C/part-734/section-734.15

International Organization for Standardization, ISO/IEC 42001:2023, Information technology, Artificial intelligence, Management system.
https://www.iso.org/standard/81230.html

National Institute of Standards and Technology, Security and Privacy Controls for Information Systems and Organizations, NIST SP 800-53 Revision 5, September 2020.
https://csrc.nist.gov/pubs/sp/800/53/r5/upd1/final

David W. Koran, AI in Defense Manufacturing: A Management Framework for Governed AI Deployment Under CMMC, DFARS, and Export Controls, August 2026.
https://davidkoran.com/ai-governance/white-papers/ai-in-defense-manufacturing/

David W. Koran, How to Build an AI Governance Program: A Records-Based Guide for Small and Midsize Organizations That Use AI, August 2026.
https://davidkoran.com/ai-governance/white-papers/how-to-build-an-ai-governance-program/