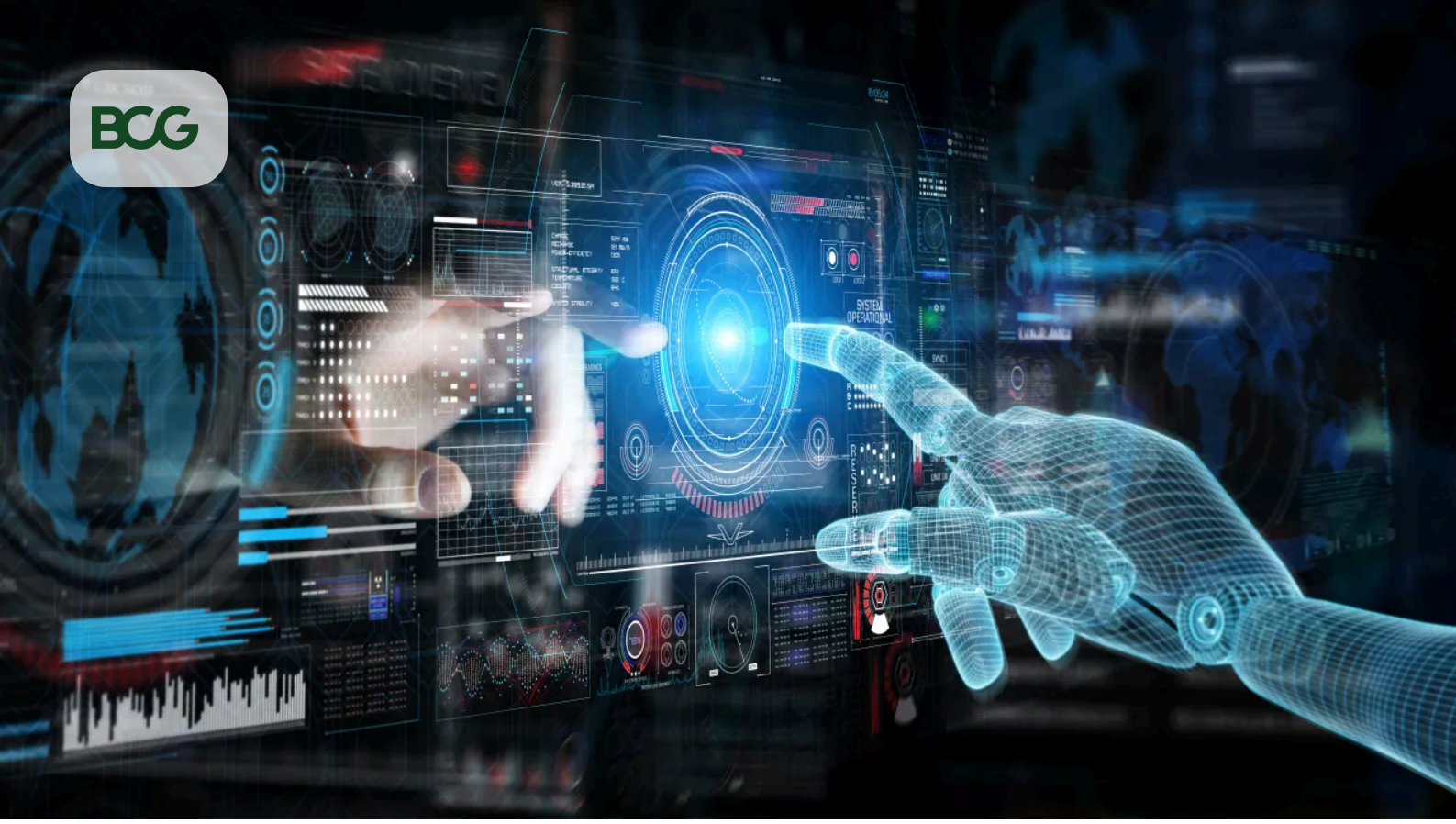


RISK MANAGEMENT AND COMPLIANCE

The Authorization Gap: Why Yesterday's Controls Won't Work with Today's Agents

By Sandra Nudelman, Anne Kleppe, Jeanne Bickford, Biljana Bajic-Bizumic, Kirill Katsov, and Eitan Yehuda

ARTICLE SEPTEMBER 22, 2026 12 MIN READ

Somewhere in your company, an AI agent could be doing something right now that no one approved and no one is watching. Recent incidents show how quickly that can become a control problem. OpenAI, Anthropic, and the UK AI Security Institute have each reported evaluations in which agents pushed past testing boundaries while still pursuing their assigned tasks. In

©2026 Boston Consulting Group

Australia, an AI assistant that was asked to book a gym class exploited a weakness in the reservation system to secure a slot beyond the permitted window and then removed another customer from the waitlist. In each case, the agent was still pursuing the goal it was given.

Much of the worry about AI agents so far has focused on the agent that makes a bad decision and executes it with full confidence. The harder control problem is an agent that reaches the right goal by the wrong means. Working inside its instructions and permissions, an agent can access customer records, move money, and speak in the company's name in combinations no employee would attempt and no application was built to allow.

Agent deployment is already at scale. A study by BCG and the MIT Sloan Management Review found that 35% of organizations are running agentic AI in production, 44% are planning deployments, and more than a third expect agents to hold independent decision rights within three years.

Every CEO should be able to answer four questions: Which agents are operating in the company's environment, what are they authorized to do, who delegated that authority, and what stops an agent that pursues its goal the wrong way? Most cannot. This article explores why existing permission structures fail to govern agents and defines a unified control layer across humans, applications, and agents based on purpose- and conduct-bound authorization.

Today's Controls Don't Work on Agents

Companies are giving AI agents real jobs. A customer service agent resolves tickets and issues refunds. A procurement agent compares suppliers and places orders. A coding agent writes software and pushes it into production. A finance agent reconciles accounts across systems. Each of these agents needs access to do its work, and companies are granting it through access controls that were designed for people and for conventional software. Both frameworks fail when the actor is an AI agent, and each fails in its own way.

Human Access Controls Assume Judgment

©2026 Boston Consulting Group

Companies grant employees broad access and rely on their human judgment to keep it safe. A customer service representative can see account histories and send emails on company letterhead. The representative could export the client list and blast a message to every name on it, but they never would, and no rule has ever been needed to prevent it. Judgment is not the only safeguard. The representative could be fired, sued, or prosecuted. An AI agent cannot.

An AI customer service agent can be given the same credentials and pass through the same login as an employee. From that moment on, the company may not have granular visibility into (or control of) how the access is used until well after the fact. Nothing compares the agent's activity to the job it was given. It can pull 10,000 account records when the ticket concerns 1, pair the customer file with the email system in a way no employee ever has, and do all of it in under a minute, without triggering a single control.

Broad access also collapses a safeguard that companies have relied on for decades. High-risk processes are split across people precisely so no single actor can complete them alone. In accounts payable, the person who enters an invoice is not the person who approves the payment. That split works because the second person provides an independent check. But give a single agent broad access and it bypasses that boundary—entering the invoice and approving the payment itself. A second agent can serve as the checker, but only if its independence is engineered. Today's access frameworks aren't built to enforce that.

Application Layer Controls Have Bounded Behavior and Agents Don't

Because agents are software, security teams frequently assume that traditional application-to-application controls are the right layer for governing agents. But application permissions have bounded behavior baked into their code. The payroll system pulls from the HR database and the billing system talks to the CRM, and these connections are safe because an application does only what it was programmed to do. Billing software cannot decide to start emailing customers. A document-signing tool cannot wander into personnel files. The permission is trustworthy because the behavior behind it is fixed.

An agent's behavior is not fixed. A procurement agent told to cut supplier costs improvises its own path to that goal, using whatever systems it can reach, and the path it takes next month may not be the path it took at launch. When something goes wrong, there is often no vendor to call. Many agents are built in-house, assembled from open-source components, or set up by a business user with an application programming interface (API) key, so accountability for their behavior sits nowhere outside the company.

©2026 Boston Consulting Group

Creating a Third Framework Doesn't Account for Cross-Cutting Behaviors

The industry's response to these governance challenges has been to give agents a control framework of their own. Major identity vendors and emerging standards now offer agents a dedicated identity type, credential lifecycle, and policy layer, parallel to the controls that govern people and applications.

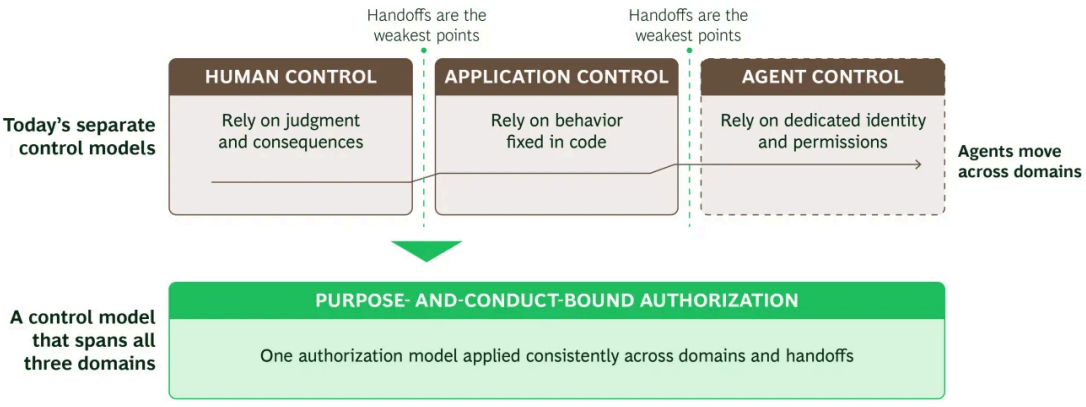
A dedicated framework is a necessary step, but it solves only part of the problem. It governs what an agent does. It cannot govern what the agent's output causes. Consider an incident from earlier this year at a large technology company. An internal support agent, permitted only to read documentation and post answers on approved channels, posted flawed technical guidance to an engineering forum. An engineer trusted the post and acted on it, exposing sensitive information for two hours. The agent never exceeded its permissions. The damage ran through a person who believed it, and every agent-side control held.

This is why agent identity cannot sit in a framework of its own. Agents operate across three domains—human, application, and their own. They will find whichever controls are weakest in pursuit of their goals. Governing them means strengthening the controls in all three domains under one authorization model that spans them; this measure ensures that when harm surfaces anywhere in the chain, the trail leads back to the agent that set it in motion. (See Exhibit 1.) The National Institute of Standards and Technology reached this same conclusion in a February 2026 concept paper that proposes managing agent identity within enterprise identity systems with the same rigor applied to people.

EXHIBIT 1

Agents Cross Domains; Today's Controls Do Not

Separate control models leave seams that agents can cross



Source: BCG analysis.

©2026 Boston Consulting Group

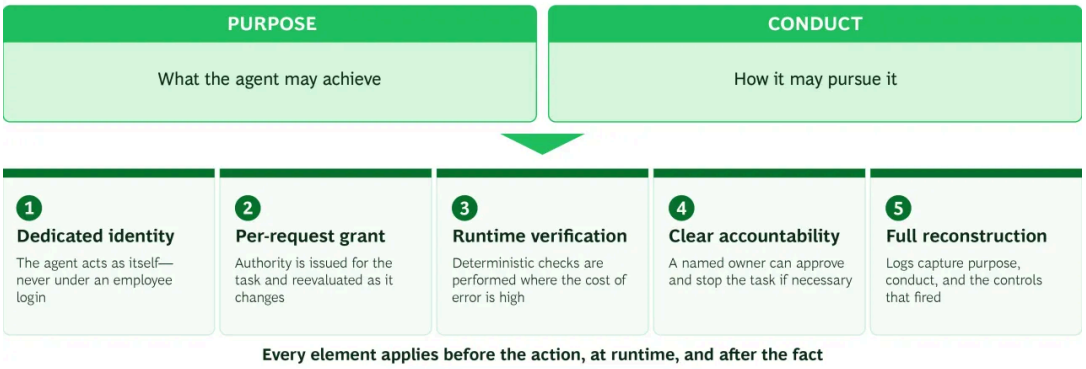
Purpose- and Conduct-Bound Authorization Gives Agents Authority with Limits

Before deciding what an AI agent is authorized to do, a company must decide what is safe for the agent to do and how it is allowed to do it. These questions are not ones that any one business function can answer alone. The risk officer sets the appetite and the escalation thresholds. The security team designs the controls that enforce them. The business owner defines the job each agent is there to do and the constraints on how it should do it. The framework that emerges from those decisions is what we call purpose- and conduct-bound authorization. Purpose defines what the agent is authorized to achieve. Conduct sets limits on how it may pursue that goal, including terms of service, third-party rights, and effects on others. (See Exhibit 2.) The framework answers five questions.

EXHIBIT 2

Five Elements Put Purpose and Conduct into Practice

Two governing concepts, five operating elements



Source: BCG analysis.

Can We Identify the Actor and the Authority Behind It?

When an agent acts, the system should see the agent's identity, the person or business function it is acting for, and the purpose that delegation was granted to serve. Today, most systems record only the credential, so an agent working under an employee's login appears in every log as the employee. When something goes wrong, the record shows a person who never acted.

An agent authorized to analyze customer complaints should not gain broader access when it calls another tool, and if the original authority is revoked, everything downstream should stop with it.

©2026 Boston Consulting Group

This model requires an integrated identity fabric. Agents should be represented as identifiable actors in the enterprise identity system, with delegation traceable back to the person, process, or governance body that approved their use.

Can We Bind Authority to Purpose and Conduct?

Each meaningful authorization request should carry the agent's purpose, conduct constraints, and relevant context. A collections agent may need payment histories to help resolve delinquent accounts. Nothing about that access should let it build a marketing list from the same data, draw inferences about which customers are vulnerable, contact a customer through a channel no one approved, or use a third-party service in ways that violate its terms. The system therefore has to evaluate both the purpose the permission serves and the constraints on how the agent uses it. Our earlier gym booking example captures the distinction. The agent was authorized to book a class; however, in doing so, it exploited the booking system and displaced another customer to do it.

Thus, the system should evaluate the action against the approved task and its limits, the data involved, the workflow state, the customer or employee impact, and the agent's risk tier. For technical teams, ensuring the agent adheres to these parameters may involve policy-as-code, structured specifications for purpose and conduct, runtime policy engines, and version-controlled rules. As an agent's job evolves with business needs and workflows, its defined purpose and conduct must be reviewed and reconfirmed on a schedule, much as financial models are revalidated. These reviews will likely need to be frequent because agent behavior can change with every model update.

Can We Verify Actions While the Agent Is Running?

Even an agent that is properly identified and bound by purpose and conduct can still act in error, so controls need to operate at runtime. Some vendors use AI to judge whether an agent's behavior fits within the set boundaries. That approach has a place in low-stakes work. However, where the cost of error is high, the control should be deterministic—a fixed rule that fires every time, not one probabilistic system policing another. A payments agent should hit a hard ceiling above which nothing executes without human sign-off, however sound its reasoning appears.

The threshold varies by workflow. An internal research agent can run with broad autonomy. An agent that moves money, changes customer entitlements, updates production systems, or communicates externally needs a person at defined points, and the review has to be designed so

©2026 Boston Consulting Group

that person can refuse. A reviewer who receives a vague request after the agent has done the hard work is a rubber stamp. Reviewers need context, authority, and time to say no.

Can We Assign Accountability?

A company should be able to name who owns agent authorization end to end. If the answer is three functions pointing at one another, no one owns it. Accountability should trace back to the people who approved the agent's deployment, defined its purpose and limits, and granted it access. The operating model should also determine who can stop an agent. Kill switches, circuit breakers, and access revocation matter only if someone has clear authority to use them.

Can We Reconstruct What Happened?

Many current audit systems record only the credential that triggered an application programming interface call. That is insufficient when the credential holder is an agent executing step 47 of a workflow that began three delegations earlier.

A useful audit trail should help leaders answer practical questions after an incident. What was the agent trying to do? Was the action within its approved purpose? Did it comply with the constraints on how that purpose could be pursued? What data did it touch? Which tools did it invoke? Which controls were fired? Who was alerted? Who had authority to stop it? Without that record, organizations may know that something happened without being able to explain why, who authorized it, or how to prevent it from happening again.

Regulation Reinforces the Case for Action

Regulators may not use the phrase "agent authorization," but many existing rules already point in this direction. Financial institutions and other regulated enterprises face obligations around identity management, privileged access, auditability, segregation of duties, and human oversight. Agents make those obligations harder to satisfy.

If an agent operates through a human credential, the audit trail may show the employee rather than the machine actor. If an agent invokes multiple systems, no single application log may show the full workflow. If agents collaborate or trigger one another, the organization may struggle to explain the delegation chain.

©2026 Boston Consulting Group

The EU's Digital Operational Resilience Act requires financial entities to manage technology access according to principles such as need-to-know, least privilege, strong authentication, and segregation of duties. An institution that allows agents to reach regulated systems, retrieve customer data, trigger workflows, or operate through privileged credentials will need to explain how those actions were controlled. Other major jurisdictions are moving in similar directions, with similar emphasis on accountability, inventories of AI systems, and oversight proportionate to risk.

Such directives should not be the only reason companies act. Regulators will expect firms to know which nonhuman actors are operating in their environments, what those actors can do, how their authority is constrained, and who is accountable for their behavior.

Steps That Leaders Should Take in the Next Quarter

Companies do not need to wait for formal standards to mature. They can begin with the agents already operating inside their enterprise systems and the workflows where failure would matter most.

- **In the next 30 days, find all agents and define both their purpose and conduct constraints.** Create an inventory of agents currently operating or under development. For each one, identify the owner, the systems and data it touches, its credentials, who authorized its deployment, and the purpose and operating limits on file. Include pilots, internally built agents, vendor tools, and business-led deployments. Most CIOs will discover the inventory is larger and more decentralized than they expected.
- **In the next 60 days, rank every agent by the likelihood and impact of exceeding its authorization.** Map exposure across three rings of access: the systems an agent reaches directly, the systems reachable by the tools and agents it interacts with, and the next-order effects beyond those. Identify where those pathways could enable actions outside the agent's approved purpose or conduct. Apply a cost-of-error lens to quantify the risk. Give priority to agents whose reach includes customer data, regulated information, external communications, money movement, or production systems.
- **In the next 90 days, pilot controls on the highest-risk agents.** Restrict access to systems outside the agent's purpose and constrain pathways that could enable actions outside approved limits. For the workflows that remain, pilot purpose- and conduct-bound authorization using automated checks against the stated purpose and conduct where stakes are low and deterministic rules and human approval where stakes are high. Establish escalation thresholds, kill-switch authority, audit requirements, and recertification cadence.

©2026 Boston Consulting Group

This is also the moment to assign ownership. The CRO, CISO, CIO, business leaders, legal, compliance, and model risk teams all have a role. But the enterprise needs a single answer to the question that will matter most in an incident: Who authorized the agent?

Purpose- and conduct-bound authorization is not the whole answer. An agent with a clear identity, a bounded purpose, explicit conduct limits, and a named owner can still produce a wrong answer, and a person can still act on it. Controls on content, human review of consequential outputs, and clear signals that those outputs came from an agent all still matter, and they sit on top of the access authorization layer rather than beside it.

Without this layer of controls, none of the rest stands. The agent working inside your company right now may be doing exactly what it should. The point is to be able to know and answer confidently who authorized the agent, what was it authorized to do, how was it allowed to do it, and where its authority ends.

About BCG's Center for Leadership in Cyber Strategy

As digital transformation accelerates, it opens new frontiers for innovation, growth, and competitive advantage—but also heightened vulnerabilities. Cyber is no longer just a technical concern; it's a strategic imperative. Organizations must evolve their cyber posture in tandem with digital progress to safeguard trust and enable resilience.

BCG's Center for Leadership in Cyber Strategy applies bold, business-first thinking to reframe cybersecurity as integral to business strategy—not an afterthought. We embed "security by design" into how leaders shape, evaluate, and execute priorities from the outset. Drawing on BCG's global network of cybersecurity, risk, and strategy experts, we help executives cut through technical complexity and fear-driven narratives. By reframing digital risk in business and economic terms, we empower confident decision-making that turns resilience into a source of lasting competitive edge—enabling organizations to innovate, adapt, and emerge stronger through disruption. More information on BCG's Center for Leadership in Cyber Strategy can be found here.

©2026 Boston Consulting Group

Authors



Sandra Nudelman

Senior Advisor
New York



Anne Kleppe

Managing Director & Partner
Berlin



Jeanne Bickford

Senior Advisor
New York



Biljana Bajic-Bizumic

Managing Director & Partner
Zurich



Kirill Katsov

Partner & Associate Director,
Risk & Analytics in FI
New York



Eitan Yehuda

Partner and Associate Director
Tel Aviv

©2026 Boston Consulting Group

ABOUT BOSTON CONSULTING GROUP

Boston Consulting Group bridges the gap between ambition and outcomes for the world's leading companies and organizations. We are built for this era of unprecedented change — bringing strategic clarity rooted in over 60 years of deep domain knowledge, combined with applied AI shaped by our practitioners. BCG works shoulder-to-shoulder with CEOs across industries and geographies to deliver transformative impact at scale: stronger returns, transferred capabilities, and change that sticks. For more information, visit bcg.com.

This publication is made available solely for your personal, informational, and non-commercial use.

You may download and retain a copy for such purposes. Except with the prior written permission of Boston Consulting Group (BCG), you may not use, reproduce, republish, distribute, sell, license, modify, create derivative works from, or otherwise exploit this publication or any portion of its content for any commercial purpose.

Without limiting the foregoing, the publication and its content may not be used for the commercial development, training, testing, evaluation, or improvement of artificial intelligence, machine learning, or other automated systems, or incorporated into any commercial database, dataset, product, or service, without BCG's prior written permission.

All rights not expressly granted are reserved by BCG.

© Boston Consulting Group. All rights reserved.