

JUNE 22, 2026 | PUBLICATION

Client Alert: 9 Insidious Ways that AI Agents Increase Your Risk

What happens when you take all the potential problems inherent in the first generation of commercially available generative artificial intelligence (AI) – hallucinations, overconfidence, sycophancy, intellectual property (IP) theft, and mistaken exposure of confidential information – and add authority and autonomy?

Agentic AI.

These are more than just basic AI chatbots or document retrieval agents. These are systems designed to manage your tasks for you, including making all the decisions you would ordinarily make along the way.

OpenClaw can manage email, calendars, files, browser sessions, and other business tools. Claude Code can read source code, execute shell commands, modify repositories, and interact with development systems. Mythos-class cybersecurity models can discover software vulnerabilities, validate flaws, and (theoretically, at least) accelerate remediation. Similar capabilities continue to appear across the AI market.

It's very cool. It's what Jarvis did for Iron Man, and if you only look at the positives, it certainly appears to be the dawn of a bright new era in personal AI assistants. If you ignore the risks.

Those risks are often technical at the start. The legal consequences come later, but as we've seen, they can include privacy violations, breach-notification duties, contract claims, employment disputes, regulatory enforcement, IP loss, securities disclosures, fraud, and negligence. All because you delegated your tasks to an AI agent.

Here's the key point: the rules still apply. You and your business are still responsible for the tasks you delegate. Delegating to an AI agent does not exempt you from responsibility. The agent can perform the work, but the organization remains accountable for the outcome.

Here are nine major business risks you face in the new age of Agentic AI:

INDUSTRY SECTOR

Technology

SERVICE LINE

Data Privacy
Technology

RELATED PROFESSIONALS

Brian C. Focht

MEDIA CONTACT

Wendy M. Byrne
wbyrne@shumaker.com

1. Unauthorized Business Actions via Indirect Prompt Injection

One key weakness of AI agents is that they operate based on instructions. Prompt Injection is a type of malicious attack that improperly gives your AI agent malicious instructions, which your agent then acts on.

Indirect prompt injection places those instructions inside the content processed by the agent.

The malicious instruction may appear in an email, on a website, in a document, in a source code comment, in a calendar invitation, in a customer support ticket, or in a tool response. The agent reads the content as part of an ordinary task. The agent then treats the attacker's instruction as part of the assignment.

A chatbot may produce a bad answer. An agent may disclose information, change a record, approve a request, send a message, or execute code.

The legal implications depend on what action your agent takes. Disclosure of personal information may trigger privacy laws, contractual notice duties, or breach-notification requirements. Disclosure of confidential business information may destroy trade-secret protection. A fraudulent payment may create disputes among the business, the bank, the insurer, the vendor, and the customer (oh, and it will likely cost you a lot of money).

You may also face allegations of negligence, arguing that the company gave an agent broad access without adequate testing, monitoring, or transaction controls.

What can you do?

Limit each agent to approved sources, approved destinations, and a narrow set of actions. Require separate human authorization for external disclosures, payments, code changes, access grants, and other high-impact decisions. Add prompt-injection testing to vendor reviews, privacy assessments, incident-response exercises, and written AI risk assessments.

2. A Small Mistake Becomes a Major Catastrophe

AI agents become useful through access. The more access, the more useful, but increasing access also increases legal risks.

An agent may receive permission to read email, access cloud storage, use Application Programming Interface (API) keys, modify source code, open customer records, or connect to production systems. A single agent may inherit much of a user's digital authority.

Microsoft has advised organizations to treat self-hosted OpenClaw deployments as untrusted code execution with persistent credentials. Similar concerns apply to coding agents, browser agents, and systems connected to enterprise software.

Broad access can turn a limited incident into a reportable breach. The compromised agent may reach regulated personal information, privileged communications, trade secrets, payment data, health information, or confidential customer material.

Incidents lead to investigations. Regulators and class action plaintiffs often evaluate the reasonableness of your access controls during those investigations to decide whether and to what extent your business should be penalized. Cyber insurers may also examine privileged access, credential management, segmentation, and multifactor authentication during coverage disputes.

What can you do?

Give every agent a separate business identity with narrow permissions, short-lived credentials, and complete activity logs. Block routine access to production systems, privileged communications, regulated information, and bulk customer records. Require additional approval for deletion, data export, credential use, permission changes, and transactions above established limits.

Convenience provides little legal value during an investigation.

3. Errors, Secrets, and Malicious Instructions Kept for Years (in Discoverable Format)

Memory improves agent performance. It also creates a new data repository.

An agent may retain user preferences, prior instructions, customer details, project history, credentials, tool results, and internal decisions. The stored information may influence future tasks, and is thus important to retain.

On the other hand, the new data repository is also a new attack surface. A poisoned memory can create a persistent security compromise. An inaccurate memory can create repeated errors. Sensitive memory can create privacy and litigation issues.

Businesses need to know whether agent memory contains personal information. Privacy rights may apply. Retention schedules may apply. Legal holds may apply. Discovery obligations may apply. The memory may also contain attorney-client communications, work product, or trade secrets. The memory may even contain information protected by contract as confidential information.

What can you do?

Define the information eligible for storage in agent memory, and prohibit the storage of credentials, privileged communications, regulated data, and unnecessary personal information. Apply retention periods, access controls, deletion procedures, legal-hold processes, and source labels to every memory store. Require review before permanently storing information derived from emails, websites, customer submissions, or other untrusted sources.

A new chat window does not erase a persistent record.

4. Trusting the Wrong Person; Acting Without Authority

AI agents are gullible, and they can fall for a lot that even the most apathetic employee wouldn't.

An attacker may impersonate an executive, customer, vendor, employee, or project owner. The message may request access, data, payment, or urgent action. The agent may rely on the displayed name, email language, title, or apparent urgency.

The resulting legal problem may extend beyond cybersecurity.

What can you do?

Businesses should define agent authority in policy and system controls. High-risk actions need identity verification through trusted records. An email signature should not establish authority. A claimed title should not establish authority.

Require agents to verify identity and authority through trusted company systems rather than names, titles, email signatures, or statements in messages.

5. Acting Without Authority

An agent may make a representation on behalf of the company. It may accept terms, authorize a refund, place an order, release information, or approve a transaction. A customer or vendor may argue that the agent possessed actual or apparent authority.

Internal instructions may not protect the company. A third party may reasonably view the agent as an authorized company representative.

What can you do?

Contracts, payment instructions, access grants, and external statements require stronger controls. The business should also preserve logs showing the identity, request, approval path, and final action.

6. Approval Without Actual Review

Many agent systems rely on human confirmation. The interface presents a summary. The user clicks approve.

This process sounds responsible. But you wouldn't say that after receiving 45 prompts to approve an action over 10 minutes.

By then, you'd do anything to make them stop. Even clicking "approve" on a request you know isn't quite right. But what about requests that are written by the very AI agent seeking permission? What happens when the AI agent determines what the request says?

Your AI agent often generates the approval request from the same context that produced the risky action. A compromised agent may describe a harmful action in harmless terms. The user approves the description and never sees the real command, recipient, destination, or data transfer.

A regulator or plaintiff may challenge the business's control process. The business may claim that a human remained in the loop. The evidence may show only a vague approval box and a routine click.

What can you do?

Meaningful approval requires verified facts. The interface should display the exact transaction, recipient, file, command, system, permission change, and expected effect. High-impact actions need stronger authentication and independent confirmation.

Human review should match the risk. Routine approval fatigue can turn a governance program into theater.

The organization needs evidence of actual oversight, not merely evidence of a button.

7. Agents Can Be Hacked (a.k.a. "The Old Fashioned Way")

Never forget, AI agents are little more than software applications. The software can contain vulnerabilities.

In the past few months, OpenClaw has patched flaws involving token theft, remote code execution, and privilege escalation. Claude Code has published advisories involving unauthorized file access, network

exfiltration, unsafe file writes, path traversal, and code execution.

Model guardrails are great, but what happens when a malicious actor can move them?

A preventable software flaw may create regulatory exposure under state reasonable-security laws, Federal Trade Commission standards, contractual security clauses, or sector-specific rules. You may also face claims based on delayed patching, inadequate configuration, or poor monitoring.

Public companies face an additional issue. A material incident may require escalation, disclosure analysis, and preservation of evidence. Statements about the company's security program must remain accurate.

What can you do?

Agent systems need ordinary security controls.

Place agents within the company's existing asset inventory, patch-management program, vulnerability-scanning process, and incident-response plan. Monitor vendor advisories, install critical fixes promptly, and document any accepted delay or compensating control. Use network separation, endpoint monitoring, secure configurations, and limited credentials to reduce the impact of a software flaw.

AI does not make basic security obsolete. It makes basic security more urgent.

8. Small Errors Spread Before Anyone Notices

"Never attribute to malice that which can be explained by stupidity." – Hanlon's Razor.

An AI agent does not need an attacker to cause serious harm.

The agent may misunderstand a goal, select the wrong tool, repeat a faulty action, or optimize the wrong result. A coding agent may alter hundreds of files. A financial agent may repeat a transaction. A customer-service agent may make the same false representation to many customers.

The same mistakes humans make all the time. Except repeated. Over and over. At scale.

Scale changes the legal consequences.

A single error may become a systemic contract breach. A mistaken customer statement may become a deceptive practices claim. A flawed employment recommendation may affect an entire applicant pool. A bad software change may cause downtime, data loss, or product liability issues. All before anyone has had a chance to detect the original problem.

Leadership and boards also face governance questions. The inquiry may focus on risk identification, approval, monitoring, escalation, and response. A company without an inventory of autonomous systems may struggle to demonstrate reasonable oversight.

What can you do?

Set limits on transaction volume, spending, file changes, messages, account updates, and other repeated actions. Test agents in isolated environments and use staged deployment prior to access to customers, production data, or business-critical systems. Maintain backups, rollback tools, shutdown procedures, and a named employee with authority to suspend the agent immediately.

Autonomy should never mean the absence of accountability.

9. Detecting Vulnerabilities Faster Creates (a Lot) More Work

Mythos and similar cybersecurity agents can identify large numbers of software vulnerabilities. Codex Security and other tools can create threat models, validate findings, produce proof-of-concept evidence, and propose patches.

This capability can improve defensive security. It can also create legal and operational pressure.

A company may discover more vulnerabilities than it can patch or be flooded with vulnerabilities that must be examined to determine their severity level, creating a logjam. Knowledge of a serious flaw can affect negligence analysis, customer notices, contractual duties, risk disclosures, and regulatory obligations. Internal reports may become evidence during later litigation.

The method of testing also matters. Vulnerability research outside authorized systems may raise computer access concerns. Distribution of exploit details may increase risk. Disclosure to maintainers, customers, government agencies, or the public requires planning and potentially a public relations team.

Automated patches introduce another problem. A proposed fix may create a new defect. Businesses still need testing, approval, and deployment controls. And we can't ignore that a vibe-coded patch to an AI-detected vulnerability all requires human review to make sure the problem hasn't just been made worse.

What can you do?

Create a written process for validating, ranking, documenting, and remediating vulnerabilities found by AI systems. Define authorization limits for testing and establish procedures for disclosure to vendors, customers, regulators, insurers, and affected partners. Require human review of proposed patches, preserve evidence of remediation decisions, and track each serious finding through verified closure.

The agent can compress discovery time. It cannot eliminate the legal duties connected to authorization, documentation, remediation, disclosure, and verification.

(BONUS) 10. Government Shuts Off Your Model on a Friday Afternoon

A business may test an AI model, negotiate a contract, train employees, and build important workflows around the service. Government action can still remove access with little warning.

The Anthropic dispute made that risk concrete. On June 12, 2026, the U.S. government directed Anthropic to suspend access to Fable 5 and Mythos 5 for foreign nationals inside and outside the United States. Anthropic stated that compliance required a broader shutdown of both models for all customers. Other Claude models remained available.

The government did not expressly order a worldwide shutdown. Anthropic made that decision, although they likely had no choice. The distinction provides little comfort to a customer who suddenly loses access.

Export controls, sanctions, licensing requirements, court orders, and national-security measures can disrupt an AI service. A U.S. company may still face the impact. Foreign employees, contractors, affiliates, support personnel, or data centers can affect the vendor's compliance obligations.

Loss of access can create contractual and regulatory consequences. A company may miss deadlines, violate

service commitments, interrupt a regulated process, or lose access to model-specific work history. A replacement model may also require new privacy, security, performance, and legal reviews.

The vendor agreement may offer limited protection. Government orders often fall under suspension, compliance, or force majeure provisions. Service-level commitments may exclude those events. A refund of prepaid fees will rarely cover business interruption or customer claims.

What can you do?

Businesses should identify all important workflows tied to a given model and maintain a tested alternative. Contracts should address notice, data export, transition support, model substitution, refunds, and access to logs or work product. Companies should also monitor export restrictions and document which employees, contractors, affiliates, and locations may use the service.

An AI model should not become critical infrastructure without a practical plan for the day it disappears.

Feeling Overwhelmed? Focus on What You Can Fix.

Businesses should treat AI agents as governed enterprise systems.

Start with an inventory. Identify every agent, model, tool, extension, identity, credential, memory store, data source, and permitted action.

Assign a business owner and a technical owner. Define prohibited uses. Establish approval thresholds. Review privacy and contractual requirements. Test hostile inputs. Limit permissions. Preserve logs. Create shutdown and rollback procedures.

Vendor agreements should address performance, security, data use, retention, subprocessors, incident notice, audit rights, IP, indemnification, model changes, and service suspension.

Policies should reflect actual use, not the profit-maximizing fever dreams of the board or the fairy-tale land imagined by the AI vendor. Training should cover prompt injection, confidential information, identity verification, approvals, and incident reporting.

The question is not whether the agent appears intelligent. What we're talking about is more practical: What information can it access? What actions can it take? Who can influence it? What promises can it make? What evidence remains? Who receives notice? Who bears the loss?

An AI agent combines technical capability with business authority. That combination creates value.

It also creates risk.

The agent acts. The business is accountable.

For more information on how your business can address agentic AI vulnerabilities, please contact Brian Focht or another member of Shumaker's Technology, Data Privacy, Cybersecurity & AI Service Line.