The Rapid Emergence of AI Proxy Risk™

Scott Foote

Last Updated: 18 March 2026

Phenomenati Consulting
www.phenomenati.com

6 Liberty Square, #2736
Boston, MA 02109
(508) 709-7990 (office)

<this page is intentionally blank>

Contents

1 Executive Summary

From a combined CISO, DPO, and CAIO perspective, the core governance question is no longer whether AI can assist professionals. It is whether the organization should allow AI to stand in for them. I use the term proxy risk to describe the compound risk that arises when an autonomous agent begins to represent a professional's activity, communications, decisions, or recommendations to others. At that point, the system is no longer only a productivity tool. It becomes an operational proxy for authority, judgment, and accountability.

- Proxy risk arises when an AI system stops being a tool and starts being treated as the professional: speaking, recommending, deciding, or acting in ways that others reasonably rely on as the professional's own conduct. [1][3][4][5]
- The trend is accelerating because enterprise AI platforms are moving toward tool-connected, multi-agent, interoperable workflows across data, applications, and communications channels. [1][2][3][4][6]
- The main control pattern is to separate authority, keep agents least-privileged, require real human oversight for consequential uses, and preserve replayable evidence of what the agent saw, recommended, and did. [8][9][10][11][16][17][18]

2 The Shift from Assistant to Proxy

For the past few years, enterprise AI has been primarily framed mainly as an assistant/copilot drafting support, search support, or workflow acceleration. That framing is now incomplete. **OpenAI's March 2025 launch** of agent tooling defined agents as systems that "*independently accomplish tasks on behalf of users*" and paired that definition with built-in tools, an Agents SDK, and observability features for *orchestration* and *tracing*. **Anthropic's Model Context Protocol (MCP)** likewise aims to make secure, two-way connections between *enterprise data* and *AI tools* a standard pattern. [1][2]

Google's A2A protocol is designed so agents can *communicate*, *exchange* information, and *coordinate actions* across enterprise platforms, while **Agentspace** is being positioned as a unified enterprise search and agent-adoption layer. **Salesforce** describes AI agents as "***digital labor***" and markets Agentforce as autonomous software that answers *questions* and takes *actions*. **Microsoft** reports that leaders expect teams within five years to **redesign business processes with AI** (38%), build **multi-agent systems to automate complex tasks** (42%), train agents (41%), and manage them (36%). [3][4][5][6]

Once an AI system can search, retrieve, draft, hand off to another agent, update a record, or communicate externally, it is no longer merely *assisting* a professional. It is *becoming* that professional's ***operational proxy***. That is the boundary at which *model risk* becomes ***proxy risk***.

3 Why Proxy Risk is Growing Now

Three forces are converging. First, ***autonomy* is increasing**: enterprise platforms are not just exposing models, they are exposing tool use, workflows, and multi-agent orchestration. Second, ***connectivity* is increasing**: MCP and A2A are moving the industry toward standard ways for agents to connect to data, APIs, and other agents. Third, ***regulation* is hardening**: the ***EU AI Act*** rollout is already underway, with *AI literacy* and *prohibitions* applying from 2 February 2025, *GPAI rules* from 2 August 2025, and most of the Act... including Article 50 *transparency* obligations... applying from 2 August 2026. [1][2][3][7][8]

The compliance meaning is straightforward: organizations can no longer treat agentic AI as an *experimental sidecar*. The UK ICO reminds organizations that individuals have the right *not* to be subject to solely automated decisions with legal or similarly significant effects, and that human involvement must be active rather than token. For high-risk EU use cases, the AI Act also leans explicitly on record-keeping, traceability, and competent human oversight. [9][17][18]

3.1 Risk Domains at a Glance

Risk Domain	How it Manifests	Primary Control Focus
Representation	External messages or policy answers are relied on as if the professional or organization said them.	Disclosure, approval gates, content QA
Decision	A “recommendation” becomes a de facto decision because humans rarely challenge it.	Substantive review, override rights
Access & Action	Delegated permissions, tool chaining, and prompt injection expand blast radius.	Least privilege, session isolation
Trust & Identity	AI-originated messages normalize impersonation and make fraud more believable.	Provenance, out-of-band verification
Judgment	Confabulation, monoculture, drift, and overreliance reduce quality of professional judgment.	Source checking, diverse review
Accountability	No replayable evidence of what the agent saw, recommended, or changed.	Logs, observability, retention

4 The Risk Domains that are Already Manifesting

4.1 Representation Risk

When an agent communicates externally, the organization owns the statement. The clearest early case study is **Moffatt v. Air Canada**. The British Columbia Civil Resolution Tribunal held that *Air Canada was responsible for misinformation given by its chatbot*, stressing that the chatbot was still part of Air Canada's website and that it made no difference whether the information came from a static page or a chatbot. The **New York City Comptroller's** December 2025 audit of MyCity found that the chatbot appeared unable to provide accurate or consistent information; independent testing found inconsistent answers to identical questions, and 71.4% of users who submitted feedback were dissatisfied. [12][13]

4.2 Decision and Recommendation Risk

Many organizations say the AI is "*only recommending*." In practice, a **recommendation** that is *routinely accepted*, *embedded in workflow*, and *rarely challenged* becomes **a de facto decision**. The ICO's guidance is useful here: a process is not meaningfully human if the person is merely applying the system's output rather than weighing it up and exercising discretion. In other words, the label "recommendation" does not remove responsibility if human review is ceremonial. [9]

4.3 Access and Action Risk

Connected agents are classic *confused deputies*. They may inherit delegated permissions to search files, read mail, browse the web, call APIs, or write back into systems. OWASP warns that indirect prompt injections can arrive through websites or files and change an agent's behavior in unintended ways. Its 2026 MCP security guide adds that MCP servers operate with delegated user permissions, dynamic tool-based architectures, and chained tool calls, which magnify the impact of a single flaw. Once the agent has write authority, every *untrusted input* becomes a potential instruction channel. [10][11]

4.4 Trust and Impersonation Risk

Proxy risk is also an ***identity*** *problem*. The more natural and normalized AI-originated communications become, the easier it is for attackers to *impersonate* executives, regulators, clients, or the system itself. The **FBI warned** in December 2024 that criminals are using generative AI to make financial fraud more believable and scalable. In May 2025 it separately warned that malicious actors were impersonating senior US officials using text and AI-generated voice messages and advised recipients not to assume those messages were authentic. [14][15]

4.5 Judgment Risk

Not all proxy failures look like breaches or lawsuits. Some look like *eroded judgment*. **NIST's Generative AI Profile** repeatedly flags *confabulation*, information *integrity*, information *security*, and human-AI *configuration* as recurring risk categories. It recommends empirically validated capability **testing**, **review** of sources and citations, **monitoring** of human-AI configurations, and even **tracking** *anthropomorphization* in interfaces. NIST also warns about "*algorithmic monoculture*" when the same models are repeatedly used in consequential settings. [16]

4.6 Accountability Risk

The final risk is **evidentiary**. When an agent drafts, recommends, communicates, or executes, the organization must be able to **reconstruct what happened**: the model version, instructions, retrieved context, tool calls, approvals, outputs, and policy checks. **OpenAI** is already **packaging observability** as part of its agent stack, NIST emphasizes documentation and testing in deployment-like conditions, and the EU's

high-risk regime explicitly leans on *record-keeping* and *logs* to support traceability and monitoring. Without that evidence, **post-incident accountability becomes speculative**. [1][16][17][18]

> ***"Treat every agent with write permissions as a privileged service account with a mouth."***

5 Recommendations to Mitigate Proxy Risk

5.1 Govern Agents as Privileged Actors, not Product Features

Every agent needs a named **business owner**, a **defined purpose**, an **authority boundary**, approved **data sources**, approved **tools**, a **logging standard**, a **fallback path**, and a **kill switch**. The right internal analogue is *not* a chatbot widget. It is *a privileged service account* combined with *an articulate junior operator*.

5.2 Separate the Rights to Recommend, Communicate, Commit, and Execute

These are different powers and should **not be bundled** by default. Many use cases should stop at assist or draft. Fewer should be allowed to communicate externally. Far fewer should be allowed to *approve*, *transact*, or *bind* the enterprise. A practical **maturity model** or **Decision Taxonomy** is: assist → recommend → draft → act-with-approval → bounded autonomy.

5.3 Engineer for Least Privilege and Hostile Inputs

Use **read-only access by default**, **narrow scopes**, **short-lived tokens**, **per-session authorization**, and **strict allowlists** for *write* actions. **Sandbox** third-party connectors and MCP servers. Treat webpages, files, email, tickets, and chat messages as potentially *adversarial* inputs, *not* neutral context. [10][11]

5.4 Preserve Substantive Human Oversight for Consequential Uses

In employment, benefits, health, safety, credit, disciplinary, legal, or customer-commitment contexts, the **human reviewer** must be empowered to **challenge**, **override**, and **document** their ***reasoning***. The goal is *not* a rubber stamp; it is ***accountable judgment***. [9][18]

5.5 Test the Proxy, Not Just the Model

Traditional benchmark scores are not enough. Organizations need **scenario testing** for *misrepresentation*, *unauthorized* commitments, *data leakage*, *prompt injection*, *bias*, *impersonation*, and *unsafe* agent *handoffs*. NIST recommends **empirically validated testing** and **deployment-like evaluation**; OWASP emphasizes **adversarial threat thinking** for agentic systems. [10][11][16]

5.6 Make Traceability and Provenance Operational

Keep replayable logs for material actions. Retain retrieved **sources**, **tool invocations**, **model versions**, **approval steps**, and **external messages**. Add ***disclosure*** and ***provenance*** cues where AI is interacting with people or generating public-facing content. That is both a *governance* control and, increasingly, a *regulatory* expectation. [8][17][18]

5.7 Train Professionals to Verify Both the Output and the Channel

Teach staff to verify unusual requests through an independent channel, especially when the message purports to come from leadership or a trusted agent. The FBI's advice remains sound: do not assume the message is authentic, and verify identity independently. In an agentic enterprise, skepticism is not friction; it is control. [15]

6 Conclusion

The **strategic mistake** would be to *treat proxy risk as a narrow technical flaw*. It is broader than *hallucination*, broader than *privacy*, and broader than *cyber*. It is the **compound risk** created when an AI system is allowed to stand in for a professional without a matching *control system*.

The question executives should now ask is *not* "*Can the model do the task*?" but "***Should the organization let the model represent the professional at this level of authority*?**" Organizations will absolutely use agents, and they should. The productivity upside is real. But the safest and most resilient organizations will automate assistance aggressively while delegating authority cautiously.

In the years ahead, that distinction... between tool and proxy... will determine whether agentic AI becomes a **force multiplier for professional judgment** or a scalable **source of unmanaged institutional risk**.

7 References

[1] OpenAI. “New tools for building agents.” 11 Mar 2025. https://openai.com/index/new-tools-for-building-agents/

[2] Anthropic. “Introducing the Model Context Protocol.” 25 Nov 2024. https://www.anthropic.com/news/model-context-protocol

[3] Google Developers Blog. “Announcing the Agent2Agent Protocol (A2A).” 9 Apr 2025. https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/

[4] Google Cloud Blog. “Google Agentspace enables the agent-driven enterprise.” 9 Apr 2025. https://cloud.google.com/blog/products/ai-machine-learning/google-agentspace-enables-the-agent-driven-enterprise

[5] Salesforce. “What Is Digital Labor?” https://www.salesforce.com/agentforce/digital-labor/

[6] Microsoft WorkLab. “2025: The year the Frontier Firm is born.” 23 Apr 2025. https://www.microsoft.com/en-us/worklab/work-trend-index/2025-the-year-the-frontier-firm-is-born

[7] European Commission AI Act Service Desk. “Timeline for the Implementation of the EU AI Act.” https://ai-act-service-desk.ec.europa.eu/en/ai-act/timeline/timeline-implementation-eu-ai-act

[8] European Commission. “Guidelines and Code of Practice on transparent AI systems.” 26 Sep 2025. https://digital-strategy.ec.europa.eu/en/faqs/guidelines-and-code-practice-transparent-ai-systems

[9] UK Information Commissioner’s Office. “What does the UK GDPR say about automated decision-making and profiling?” https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/individual-rights/automated-decision-making-and-profiling/what-does-the-uk-gdpr-say-about-automated-decision-making-and-profiling/

[10] OWASP Gen AI Security Project. “LLM01:2025 Prompt Injection.” https://genai.owasp.org/llmrisk/llm01-prompt-injection/

[11] OWASP Gen AI Security Project. “A Practical Guide for Secure MCP Server Development.” 16 Feb 2026. https://genai.owasp.org/resource/a-practical-guide-for-secure-mcp-server-development/

[12] Moffatt v. Air Canada, 2024 BCCRT 149 (CanLII). 14 Feb 2024. https://www.canlii.org/en/bc/bccrt/doc/2024/2024bccrt149/2024bccrt149.html

[13] Office of the New York City Comptroller. “Audit Report on the New York City Office of Technology and Innovation’s MyCity System.” 30 Dec 2025. https://comptroller.nyc.gov/reports/audit-report-on-the-new-york-city-office-of-technology-and-innovations-mycity-system/

[14] FBI Internet Crime Complaint Center (IC3). “Criminals Use Generative Artificial Intelligence to Facilitate Financial Fraud.” 3 Dec 2024. https://www.ic3.gov/PSA/2024/PSA241203

[15] FBI Internet Crime Complaint Center (IC3). “Senior US Officials Impersonated in Malicious Messaging Campaign.” 15 May 2025. https://www.ic3.gov/PSA/2025/PSA250515

[16] National Institute of Standards and Technology (NIST). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1). July 2024. https://doi.org/10.6028/NIST.AI.600-1

[17] European Commission AI Act Service Desk. “Article 12: Record-keeping.” https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-12

[18] European Commission AI Act Service Desk. “Article 26: Obligations of deployers of high-risk AI systems.” https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-26