POLSINELLI

Publications

May 7, 2026 • Updates

U.S. and Allies Release “Careful Adoption” Guidance for Agentic AI

Key Takeaways

- AI is accelerating cybersecurity threats by expanding the attack surface and enabling more sophisticated, scalable attacks, even as it offers potential defensive benefits.
- Last month, the limited release of new AI systems designed for cybersecurity underscored how new and fast-emerging risks are an inherent part of AI’s potential.
- Last week, the U.S. and its allies released guidance on how AI security risks for agentic AI systems can and should be addressed within established cybersecurity frameworks.
- Industry standards for AI cybersecurity are evolving rapidly, and signals included in this guidance will shape the establishment of duties of care and legal obligations.

AI is rapidly reshaping cybersecurity risk, not just as a defensive tool, but as a force multiplier for threat actors. When AI moved from just generating outputs (generative AI) to taking actions (agentic AI), “it crossed a legal threshold” that many users of AI may not have noticed: to interpret and reason about the state of the world, make decisions and take actions.

Agentic AI (AAI) builds on generative AI by integrating with software systems to create autonomous agents that can independently reason, plan and take actions *without requiring human intervention*. AAI can create legal risks for organizations.[1] This alert highlights some of these legal risks and offers new guidance about how to address them.

The AI Debate Continues as Adversarial Capabilities Escalate

Will AI live up to its purported potential? Will AI adoption rates translate to sustainable ROI when it is embedded in workflows to accelerate contracting and risk management — and even human resources? Will AI be transformative when, according to some, it is far better than humans at some tasks and far worse at others — something known as “jagged intelligence”?

Such questions are being debated in legislative sessions, town halls, universities and institutions of higher learning and conference rooms across the country. But as debates rage on, cybercriminals, threat actors and hackers are already using “adversarial” AI —

Related People

- Romaine C. Marshall
- Matt A. Todd
- Bryce H. Bailey
- Jennifer Bauer

Related Capabilities

- Artificial Intelligence & Machine Learning
- Privacy & Cybersecurity
- Technology

with devastating effectiveness.

For example, AI loss of control (LOC) risk — when AI systems diverge from authorized constraints and cause unintended outcomes — has accelerated. Security experts are counseling that email users should assume phishing campaigns will become dramatically more effective by "mimicking the writing styles of specific executives" and refencing "real internal projects."[2]

But just as AI can introduce new risks, it can also reduce cybersecurity risks if deployed correctly.

The Risks and Benefits of Deploying AI as a Cybersecurity Tool

Last month, Anthropic previewed the release of Mythos — a new agentic AI model touted as being so effective that it was too dangerous to release publicly. This AI model has purportedly been able to identify thousands of cybersecurity vulnerabilities that were previously undetected, including vulnerabilities going back 30 years.

Mythos was rolled out to about 50 companies for testing, and Anthropic met with government, business and technology officials to discuss the extent to which Mythos can cause harm (as well as good). Already, unauthorized access to Mythos has been reported, and competing products with similar capabilities have been or will soon be released.

The downside to new tech that excels at discovering vulnerabilities is the pressure it may place on security teams — whether through robust patching and authentication, software updates and upgrades, or because they may require segregating networks. For now, Mythos and other recently released tools cannot replicate a lot of this type of work.

While the specific impact of Mythos and GPT-5.5 is still being assessed, quantum computing will certainly impact cybersecurity across the globe. Quantum computing still sounds futuristic, but cybersecurity experts are already treating it as a serious long-term threat, especially as businesses hand more sensitive data to AI systems and other always-on digital tools.

This risk is not just a future problem. Security experts have warned for years about "harvest now, decrypt later" behavior, where attackers steal encrypted information today and save it until better tools emerge to unlock it. Put simply, the more data organizations collect and rely on, the more damaging it could be if yesterday's encryption stops holding up tomorrow.

In the meantime, Mythos effectively encapsulates a pivotal question challenging organizations: how should organizations safeguard data when using AI systems for consequential decisions? As of last Friday, according to the U.S. and four of its allies — New Zealand, Australia, the United Kingdom and Canada — existing security models provide a strong foundation.

The U.S. and Allies "Careful Adoption" Guidance

Titled *Careful Adoption of Agentic AI Services*, the guidance comprises 28 pages divided into scenarios and use cases; recommended best practices; how to incorporate accepted frameworks and standards; audience callouts that distinguish AI developers, vendors and operators; and an appendix with a breakdown of cybersecurity prerequisites to consider.[3]

The guidance is clear in issuing its bottom line upfront: "the authoring agencies strongly recommend aligning agentic AI risks and mitigation strategies with your organization's

existing security model and risk posture." The purpose of the guidance is to enable organizations that design, develop and operate agentic AI "to make informed risk assessments and mitigations."

The guidance then outlines examples of AAI risks, such as privilege compromise, scope creep, identity spoofing and agent impersonation. By issuing the guidance with key U.S. allies, the guidance is consistent with the White House's AI Action Plan released last July that said the U.S. would drive adoption of AI standards throughout the world.

The 100+ recommended best practices from the guidance revolve around these AAI themes:

- **Designing secure agents:** Understanding threats, anticipating risks and integrating mitigations before development and deployment.
- **Developing secure agents:** Requiring data training approaches to models that go beyond standard practices to include tailored techniques.
- **Deploying agents securely:** Implementing high-impact security controls (e.g., updating incident response and risk assessment procedures).
- **Operating agents securely:** Applying continuous monitoring and auditing techniques to maintain awareness and ensure traceability.

AAI guidance is increasingly important for risk management. According to a recent report by Deloitte, while only 23% of companies are using AAI today, AAI is expected to be ubiquitous within the next two years, with 74% using it at least moderately, 23% using it extensively and 5% fully integrating it as a core component of operations.

Moreover, as we have previously explained, organizations interested in mitigating AI risks within their business operations should consider signals like the above guidance as industry standards that could establish duties of care and legal obligations, which could be applied to generative AI and other emerging technologies like quantum computing.

Planning for New Risks and Threats in the Age of Adversarial AI

Organizations will need to proactively assess and address AI-specific risks in their cybersecurity risk management strategies. This effort should consider both product and workflow-specific risks related to AI tools, as well as generic risks presented by malicious use of AI by threat actors, like AI-enabled cybersecurity attacks.

A recent executive order and White House materials — including President Trump's Cyber Strategy for America and National Policy Framework for AI — reinforce prior guidance from government agencies and, as discussed in a prior alert, reiterate a preference for streamlining and implementing existing approaches.

Historically, the Federal Trade Commission and the National Institute of Standards and Technology (NIST) have developed sector-specific industry standards and best practices. Prior enforcement actions — including a 2024 order — highlight the need for risk assessments, third-party and vendor oversight, and proactive monitoring and threat detection.

NIST has been similarly active in identifying AI risks and developing responsive standards, including releasing an AI monitoring report last month.[4] The report provides guidance on identifying gaps, barriers and open questions that NIST believes risk managers and IT security teams should be evaluating, which they may never had to deal with before.

We have previously noted how developing and implementing an incident response plan,

conducting periodic risk assessments and having an updated written information security program are important components to effective cybersecurity compliance. NIST has regularly released industry-led standards relating to these governance requirements as well.

Steps to Consider for Addressing AI-Driven Cybersecurity Risks

Organizations cannot continue to treat AI as an ad hoc cybersecurity threat, especially considering new and fast-emerging risks and threats. By considering recent guidance like that released by the U.S. and its allies and monitoring technological shifts, organizations can proactively assess and evaluate AI-specific risks from at least two angles:

1. Product and Workflow-Specific Risks: How are you using AI within your organization, and what risks does a particular AI deployment present?
2. Threat Actor Use of AI: What AI-enabled cybersecurity threats or malicious use of AI should your organization be aware of, and how can you proactively mitigate them?

Developing or revising a technology asset inventory or network map, as well as updating cybersecurity incident response plans, risk assessment approaches and written information security programs will remain critical tools when developing, deploying or using high-risk AI tools and systems.

For more information on AI-driven cybersecurity risks and evolving legal expectations, contact the authors or your regular Polsinelli attorney.

[1] For an excellent overview of Agentic AI's capabilities, and legal and operational risks, *see Agentic AI for General Counsels: What You Need to Know* by Bryce Bailey here (published January 22, 2026).

[2] *When AI Becomes the Attacker's Playbook: Inside the First Major ATI -Assisted Infrastructure Breach*, WebProNews, (February 22, 2026).

[3] Last week's guidance is a follow-up to that provided by the U.S. and its allies a year ago, titled *Best Practices for Securing Data Used to Train & Operate AI Systems*. That guidance included a breakdown of the AI lifecycle, risks analyses and 10 practical steps that AI system owners can consider for developing a robust foundation for securing AI data.

[4] For a discussion on NIST and its importance when it comes to developing industry standards, including the AI Risk Management Framework AI RMF) *see* here, here, and here.