QBE Insurance

Reimagining Commercial Espionage and Fraud in the Era of Agentic AI

Introduction

Agentic artificial intelligence (AI) is an expert-level force multiplier for threat actors that amplifies the risk of commercial espionage, fraud, and insider threats to organizations regardless of industry or sector. Agentic AI models can accelerate risk across existing threats, lower the barrier to entry for less sophisticated threat actors, and introduce new and fundamentally different threats for organizations.

Although agentic AI models can act proactively and autonomously at a speed and scale that humans cannot replicate, the human factor and human behavior nevertheless remain key components for attackers and defenders alike. A back-to-basics layered security approach focused on identity/access management and behavioral monitoring – supplemented by AI-enabled threat detection tools – will likely prove the best recourse to mitigate threats. Still, it is hard to overstate the potential impact of agentic AI models and the speed and scale with which motivated threat actors are able to operate, particularly as we are only at the nascent stages of this emerging and complex threat landscape.

Key takeaways

- Agentic AI models massively accelerate the speed and scale of the attack cycle, and are able to independently accomplish tasks that previously would have required a team of hackers and in a fraction of the time
- Although agentic AI models often leverage existing threats and attack pathways, they are also creating new and emerging attack vectors
- Agentic AI models also introduce new threat dynamics to include persistent autonomy, strategic-level campaigns and objectives, and the potential displacement of humans from the process entirely
- Despite this potential, humans still play an important role in terms of goal-direction and oversight at critical decision points during attacks and campaigns
- Commercial espionage and fraud actors increasingly rely on AI models in their work. A QBE survey of 400 decision makers of IT, administration or insurance in businesses with 100 to 2,000 employees in the United States found that in the past year nearly three in 10 U.S. businesses (29%) experienced at least one cyber incident where AI was believed to have been used as part of the attack.[1]

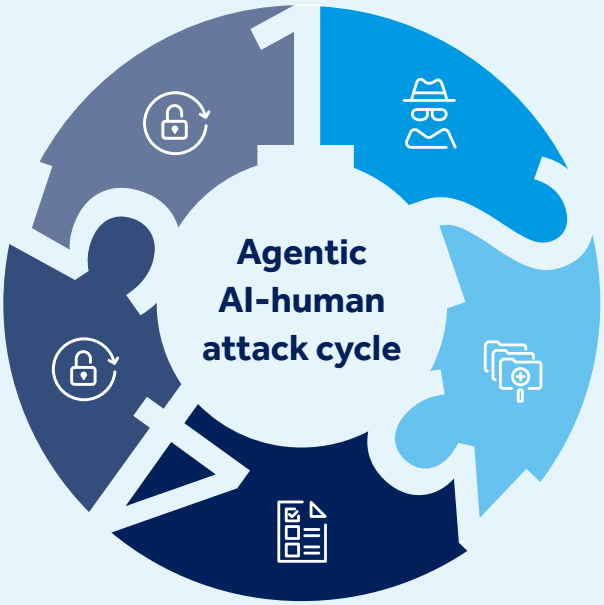
1 QBE Survey Finds Widespread Cyber Incidents and Risk Concerns Among US Businesses, QBE North America

Defining agentic AI

Agentic AI refers to "autonomous artificial intelligence systems designed to achieve complex, multi-step goals with limited human supervision," and unlike traditional generative AI, agentic AI "uses reasoning, planning, and tool orchestration to take independent action across software systems to solve problems in real-time." [2] Agentic AI models are able to act proactively and independently at times, and can engage in goal-driven behavior without human oversight or intervention. Agentic AI models can maintain long-term goals, can call application programming interfaces (APIs) and search databases, and then use the information gathered to make decisions and take further actions. They can also learn from their experiences and continuously improve, and users (including threat actors) can communicate with agents using natural language prompts, rather than having to learn a variety of new interfaces and tools.

Commercial Espionage and Fraud – Including Insiders:

Commercial espionage is "the illegal and unethical act of stealing or acquiring sensitive business information – such as trade secrets, intellectual property, or proprietary data – for the purpose of gaining a competitive advantage or financial benefit." [3] The definition of business fraud can be a bit more amorphous, but centers on intentional deception and the misrepresentation of facts or abuse of trust for financial gain. Insiders pose a potential threat in either case, as insider threats are uniquely well-placed to enable either commercial espionage or fraud on behalf of or in concert with external threat actors. In fact, agentic AI models can even act as "digital insiders" themselves, potentially causing harm unintentionally (if incorrectly trained) or deliberately if compromised by malicious threat actors.



Phase 1
Human defines target/objective

Phase 2
System gathers data/executes tasks

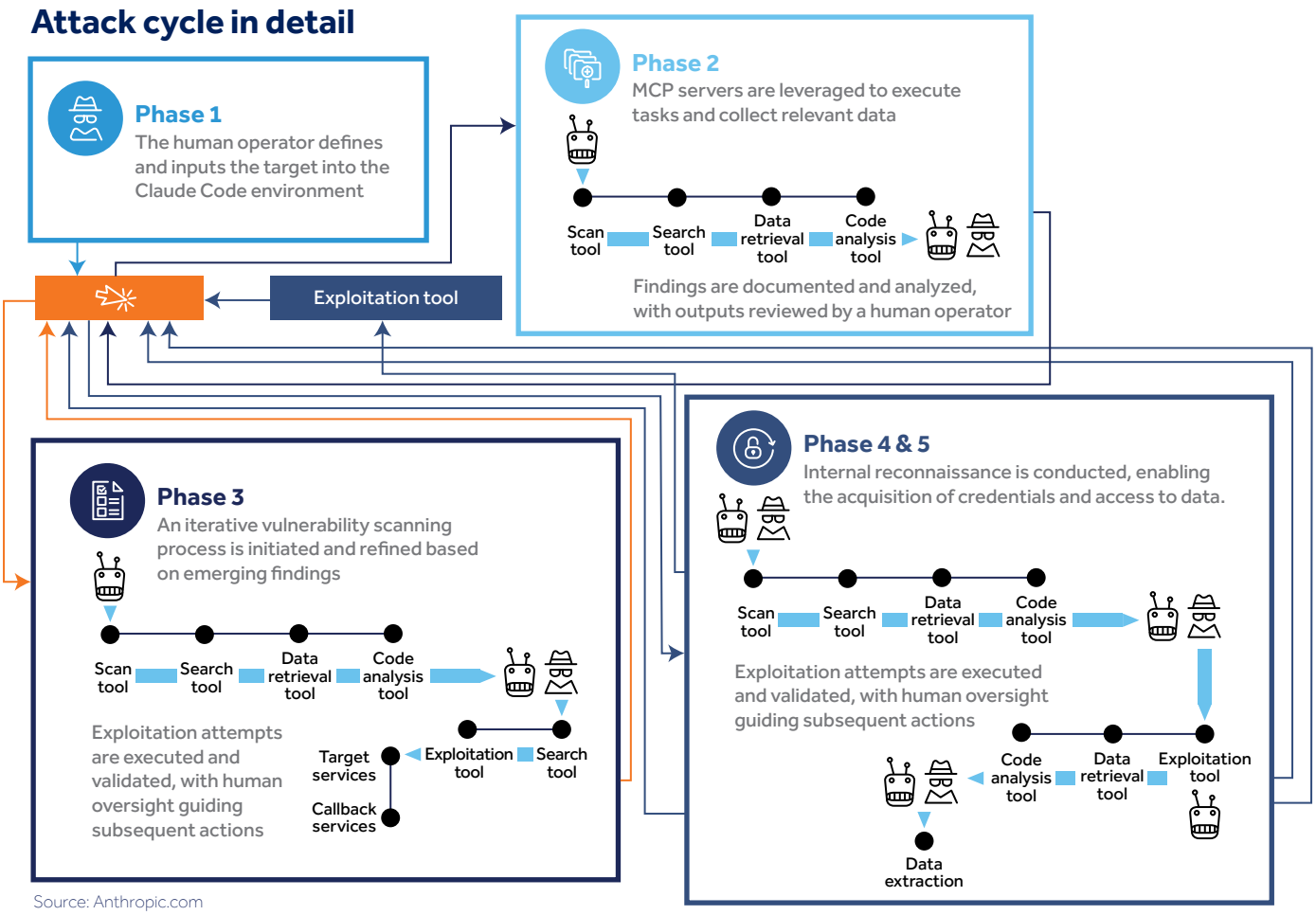
Phase 3
Actions are defined iteratively

Phase 4
Internal reconnaissance identifies access and data

Phase 5
Internal reconnaissance identifies access and data

2 "What is agentic AI?", IBM
3 "Industrial Espionage: What It Is and Why It Matters," Investopedia, Jan 29 2026

One of the most notable recent cases of agentic AI being utilized for commercial espionage purposes was detected by AI research company Anthropic in late 2025, when a suspected Chinese state-sponsored group used Anthropic's own Claude tool to attempt to breach around 30 organizations globally. The list of targets included large technology companies, financial institutions, chemical manufacturers, and government agencies, and the threat actor group was successful in some cases, according to Anthopic investigators. All told, Anthropic investigators estimated that the attackers used Claude to conduct 80%-90% of the tactical operations involved in the campaign, including for reconnaissance, to locate vulnerabilities, exfiltrate data, and create backdoors – all with little human intervention.



Source: Anthropic.com

According to Anthropic, human operators selected the initial targets for Claude, after which the AI carried out attacks using various tools (often the Model Context Protocol (MCP). Occasionally, the AI would return to its human operator (at 4-6 critical decision points per campaign) for review and further guidance . [4] Investigators at Anthropic were both impressed and alarmed at how quickly AI-driven capabilities appear to have evolved in support of such cyber activities, the scale enabled by the capabilities, and the seemingly unprecedented degree to which the human operators were able to rely on these capabilities throughout the campaign.

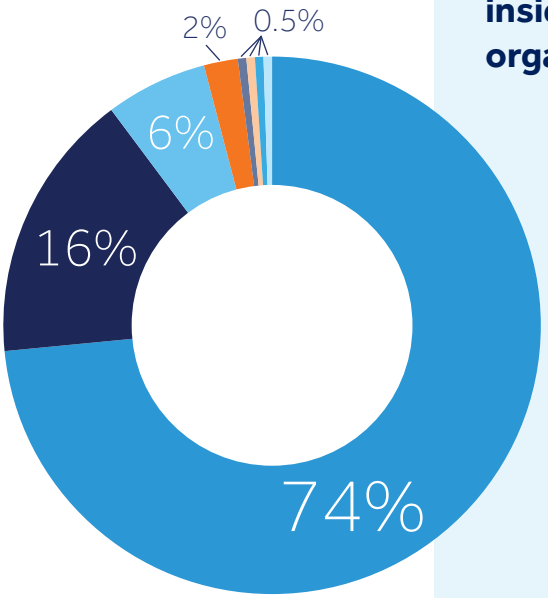
4 Disrupting the first reported AI-orchestrated cyber espionage campaign", Anthropic, Nov 13 2025

Although this campaign was ostensibly conducted by a nation-state actor, commercial espionage targeting U.S.-based companies and their intellectual property was certainly a key motivating factor. The tactics, techniques, and procedures (TTPs) employed by the most sophisticated nation-state actors are often adopted rather quickly by cybercriminals, hacktivists, and others. Interestingly, Anthropic investigators noted two potential mitigating or limiting factors involved in the campaign – factors that may or may not always be present in future attacks: First, after selecting their targets, human operators had to convince Claude to commence the attacks, and did so by "tricking it to bypass its guardrails," and instructing Claude that it was working for a legitimate cybersecurity firm engaged in "defensive testing". [5] Second, Anthropic researchers noted that Claude "occasionally hallucinated credentials or claimed to have extracted secret information which was in fact publicly-available," adding that such errors remain an inhibiting factor for such autonomous attacks. [6]

Agentic AI-enabled fraud can come in a variety of forms, including large-scale extortion, deepfake impersonation scams, and enhanced phishing and business email compromise (BEC) campaigns. In all cases, it can dramatically scale operations while lowering the barrier-to-entry for less sophisticated actors and groups.

Human fraudsters can now automate many of their criminal activities, relying on agentic AI models at each stage of their operations and even "set and forget" to a certain degree, as agentic AI can act autonomously and pivot as needed. In recent years, fraudsters initially relied on generative AI to create deepfakes and enhance social engineering attempts, a trend that continues today. These same threat actors can now use agentic AI models to manage these activities, including the creation and management of synthetic personas and the simultaneous management of interactions with targets across multiple platforms. Across the fraud landscape, agentic AI is also enabling less sophisticated fraudsters to engage in complex fraud operations at an unprecedented scale and rate.

Types of threat actors behind significant insider* incidents impacting commercial organizations, 2021-2026



- Nation-state
- Cybercriminal
- Other/unknown
- Unknown
- Coerced insider (externally motivated)
- Cyber activist (state-linked)
- Malicious insider
- State-linked proxy

Source: Control Risks

*Insider involved in the attack chain of general commercial espionage incidents, which often involve both physical and cyber

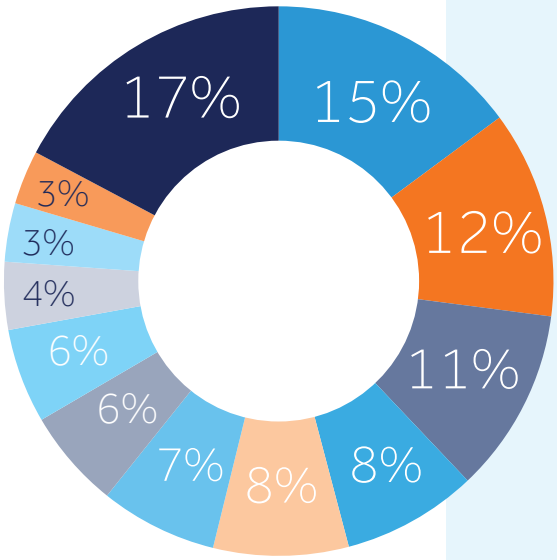
5 Disrupting the first reported AI-orchestrated cyber espionage campaign, Anthropic, Nov 13 2025
6 Disrupting the first reported AI-orchestrated cyber espionage campaign, Anthropic, Nov 13 2025

How Agentic AI Amplifies Risk

Unlike traditional AI, agentic AI models can independently navigate highly complex fraud workflows and convincingly mimic legitimate user behavior to avoid detection, while constantly adapting in real time to defensive countermeasures. Agentic AI tools can scale attacks, learn from their mistakes, retry previously unsuccessful attempts, and survive and persist in targeted systems – all without additional human intervention. The autonomy inherent in agentic AI models allows them to conduct sophisticated fraud operations such as account takeover, credential stuffing, and BEC, and at a speed and scale that humans cannot match.

In terms of commercial espionage, agentic AI models can rapidly reconnoiter for the most sensitive company databases and repositories, integrate with enterprise systems, and efficiently exfiltrate data without attracting unwanted attention or detection. Threat actors appear to view agentic AI models as "a high-speed research assistant for attack lifecycle support," and increasingly value the ability of these models to discover new vulnerabilities and generate exploits themselves, including for zero-day vulnerabilities. [7] They also view such AI systems as legitimate targets in and of themselves, particularly when those systems are integrated within target enterprise environments. At the same time, significant gaps remain in enterprise AI governance, with some estimates suggesting that less than 20% of U.S. organizations exhibit optimized AI governance frameworks – creating significant exposure to unmonitored agentic operations. [8]

General commercial espionage incidents by sector, 2021-26*



- Government and administration
- Aerospace and defence
- IT and communications
- Finance and insurance
- Manufacturing
- Energy and utilities
- Pharmaceuticals, medical supplies and health
- Education and training
- Logistics and transport
- Maritime
- Professional services
- Other

Source: Control Risks

*Incidents often involve physical and cyber

7 Google Threat Intelligence Group, "GTIG AI Threat Tracker," 11 May 2026
8 TechRT, "AI Governance Statistics 2026," 5 May 2026

The Human Factor

Even in the emerging world of agentic AI, human involvement still matters. Human operators (both attackers and defenders) play crucial roles across commercial espionage and fraud scenarios at key stages of the attack lifecycle.

On the threat actor side, humans still need to select the targets, may conduct "persona-driven jailbreaking" of the AI model (as seen in the earlier example involving Claude), develop convincing social engineering narratives, and ultimately initiate the attack. Monitoring, review, guidance and oversight still occur to varying degrees, and human attackers typically retain a guiding role at critical decision points during the campaign.

Within organizations, human operators build and train agentic AI systems, set goals, parameters, and success criteria, and provide crucial oversight. In challenging opaque situations and under intense time pressure, humans remain the final interpreters and arbiters of events and the final decision-makers. Human defenders are still responsible for deciding whether to escalate issues to executives, involve law enforcement, and, ultimately, for ensuring compliance with all applicable laws, regulations, and ethical standards.

Anthropic's Mythos and "Project Glasswing"

On April 7, 2026, Anthropic announced the launch of 'Project Glasswing', a tightly controlled, consortium based cybersecurity initiative built around leveraging Claude's 'Mythos', a nonpublic frontier AI model, to autonomously identify, chain, and validate high severity software vulnerabilities, including previously unknown zero days. Anthropic determined that Mythos is too powerful to release publicly due to its clear offensive cyber potential, and is restricting access to trusted technology, security, and critical infrastructure partners for defensive vulnerability discovery and coordinated disclosure. [8] Despite the lack of a public release for Mythos to date, developments in AI capabilities signal a general "democratization" of cyber security capabilities over time, and an increased regulatory focus for organizations – particularly for financial institutions. In the near term, accelerated vulnerability identification and disclosure are likely as Project Glasswing members apply Mythos across high-criticality software and commercially vital product bases. Threat actors will likely respond by prioritizing known-but-unpatched vulnerabilities, and accelerating credential-based social engineering, including targeting privileged accounts, CI/CD pipelines, security tooling, and agent integrations. As AI-enabled vulnerability discovery and exploit development mature, the window between disclosure and exploitation is likely to compress even further, increasing the value of compensating controls (e.g., segmentation, hardening, and egress restrictions) and continuous monitoring for anomalous behavior. Threat groups are likely to increasingly rely on living-off-the-land techniques to evade detection in cloud environments.

8 www.anthropic.com/glasswing

Near-Term Outlook

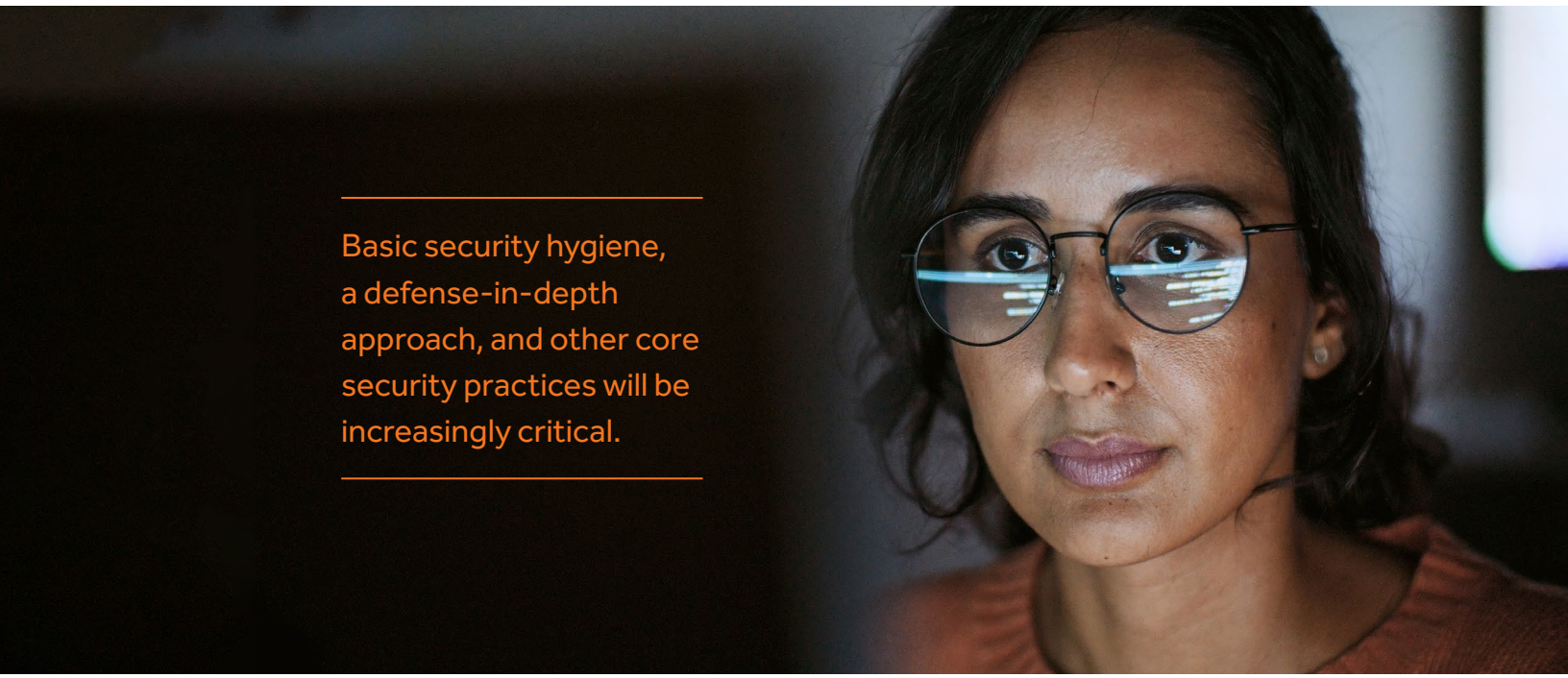
Offensive attack techniques related to ransomware deployment, phishing campaigns, and exploit development are likely to evolve rapidly, and even unsophisticated actors will leverage automated planning and tool orchestration in their work.

For organizations facing these threats, increased investment in agentic AI-enabled defensive systems and enterprise-level AI governance will need to continue apace. However, a clear divide is likely to emerge between organizations that adopt AI for defensive operation and those that are unable or unwilling to do so. In either case, regulatory scrutiny around organizational AI use will continue to intensify in the near-term, and there will be no escaping the judgement of regulatory authorities, the market, and the focused sights of malicious threat actors in this evolving threat environment.

Risk Management

Below are high-level recommendations for organizations to consider when taking steps to mitigate threats. It is not an exhaustive list, and the recommendations will vary depending on organizational circumstances.

- Basic security hygiene, a defense-in-depth approach, and other core security practices will be increasingly critical, to include enforcing strict access controls, least privilege, and continuous verification of identity and device integrity.
- Security teams will need to invest in and adopt defensive AI systems to augment the monitoring and oversight abilities of their human defenders, and will need to automate incident response playbooks and anomaly detection – as much as resources allow.



- Treat these agentic AI systems/integrations within your enterprise as privileged systems; enforce strong identity and access management (least privilege, MFA, short-lived tokens), restrict tool permissions, and implement tight network egress controls and logging for any agent that can execute code, query internal data, or interact with external services. Embed AI cybersecurity governance frameworks fully into organizational processes to monitor autonomous operations.
- Prepare for increased regulatory and customer scrutiny; ensure you can evidence patch SLAs, critical-vulnerability triage, third-party exposure management (including AI/cloud providers), and have exercised cyber crisis and communications playbooks for fast-moving, AI-accelerated incidents.
- As tools like Mythos reach red team actors, expect security windows to narrow as decision-making and security actions continue to move to the left. This means testing executive and management teams in their ability to make quick decisions, testing internal and external security teams from assumed breach scenarios to identify reaction times, and running patching cycles out of band.
- Social engineering, credential-based attacks, account compromise, man-in-the-middle, living-off-the-land and other attacks less reliant on vulnerability exploitation are likely to grow as actors move away from relying on exploit enabled attacks to automated attacks targeting human, credential, privilege and other security flaws.
- In a recent QBE survey, only two-thirds of U.S. businesses with 100 to 2,000 employees (67%) have cyber insurance, while 24% do not. Despite strong awareness of cyber risk and increased cybersecurity investment, critical gaps remain in both response readiness and insurance coverage. [9]

  Accelerating cyber risk is reshaping insurance broker conversations, shifting the focus to fast-moving, interconnected disruptions driven by digital dependencies. This dynamic is also influencing how cyber exposure is framed in conversations with customers to ensure the appropriate protection.

In the era of agentic AI, a hybrid model is emerging – with humans as supervisors, strategists, and final approvers, while AI handles operational matters and execution. For threat actors engaged in commercial espionage and fraud, this human-AI dynamic is an exponential force multiplier.

As with previous cyber crimes, sophisticated threat actors are typically early adopters, and organizations are often forced to react in their wake. While tools such as Mythos and initiatives like Project Glasswing offer significant promise for security practitioners and clear risk once similar capabilities proliferate. It is likely that attackers will once again seek to seize the early advantage in the case of agentic AI.

At this still nascent stage, adoption of agentic AI will be uneven, requiring organizations to reinforce foundational security practices while integrating agentic AI-enabled defenses to build resilience in the near term.

9 QBE Survey Finds Widespread Cyber Incidents and Risk Concerns Among US Businesses, QBE North America

This report was produced by
QBE North America and Control Risks

About QBE North America

QBE North America is a global insurance leader that gets to the heart of what's at risk for customers. Part of QBE Insurance Group Limited, QBE North America reported Gross Written Premiums in 2025 of $7.7 billion. QBE Insurance Group's results can be found at qbe.com. Headquartered in Sydney, Australia, QBE operates out of 26 countries around the globe, with a presence in every key insurance market. The North America division, headquartered in New York, conducts business primarily through its insurance company subsidiaries. The actual terms and conditions of any insurance coverage are subject to the language of the policies as issued. Additional information can be found at qbe.com/us or follow QBE North America on LinkedIn, Facebook and Instagram.

For more information, visit **qbe.com/us/cyber.**

QBE Insurance