

AI-ENABLED CYBER RISK · PRACTICAL BUSINESS GUIDE

AI-Assisted Cybercrime: What Businesses Should Watch For

A pragmatic guide to how artificial intelligence is changing established cyber risks - and the controls that still matter most.

CARRATU INTERNATIONAL / 23 AUGUST 2026



CONTENT	IDENTITY	ACCESS	DECISION
---------	----------	--------	----------

Artificial intelligence has not made conventional cyber security obsolete. It is making parts of the attack process faster, cheaper and easier to scale.

The practical challenge is therefore not identifying whether every suspicious email, voice call or piece of code was generated by AI. It is ensuring that established controls remain effective when attackers can research a target, construct a persuasive approach and adapt their methods much more quickly.

The change is acceleration, not reinvention

Most successful cyber attacks continue to depend on familiar weaknesses: compromised credentials, deceptive communications, unpatched software, excessive privileges, poorly configured systems and human error.

What AI changes is the friction around parts of that process. It can help research a target, organise information, translate technical material, produce convincing text, troubleshoot code and process large volumes of data. The UK National Cyber Security Centre assesses that AI will almost certainly make elements of cyber intrusion more effective and efficient, while the near-term effect will largely be the evolution and enhancement of existing tactics rather than wholly new threat vectors. ^[1]

OpenAI has reported a similar pattern from malicious activity it has disrupted: threat actors typically combine AI with ordinary websites, social media, infrastructure and other tools. ^[2] Microsoft describes AI as being operationalised across the attack lifecycle to reduce technical friction and shorten decision cycles while human operators retain control over targeting and deployment. ^[3]

Practical implication: do not focus on proving that an attack was “made by AI”. Ask whether your controls remain effective when research, writing, translation and adaptation can happen faster.

Where AI can assist an attacker

Stage	How AI changes the task	Defensive emphasis
Reconnaissance	Organising public information about people, suppliers, technology and business relationships.	Reduce unnecessary operational disclosure; assume public fragments can be combined quickly.
Social engineering	Drafting polished, tailored messages in multiple languages and producing many variants.	Judge the requested action, not the quality of the writing.
Coding and vulnerability research	Explaining unfamiliar software, debugging code and accelerating research into known weaknesses.	Patch promptly, especially internet-facing services and security appliances.
Post-compromise triage	Classifying files, summarising correspondence and identifying potentially valuable data.	Limit access, monitor unusual downloads and segment sensitive information.
Impersonation	Supporting convincing text, synthetic voice, images or other material around a plausible identity.	Use independent verification and robust payment controls.

Phishing is becoming harder to recognise by appearance alone

For years, staff were advised to look for poor spelling, strange grammar and awkward phrasing. Those clues can still matter, but they are no longer dependable. Generative AI can produce fluent correspondence, adapt tone and terminology, translate approaches into numerous languages and create many variants of the same message.

The safer habit is to assess the requested action. A message that asks someone to change bank details, disclose credentials, install software, bypass a normal process, approve an unexpected payment or provide sensitive information deserves verification regardless of how professional it looks.

Verification principle: confirm consequential requests through a separate, established route. Do not rely on replying to the same message or using a telephone number supplied inside it.

Public information also becomes more useful when AI can combine it quickly. Company websites, LinkedIn profiles, job advertisements, supplier announcements and social media may each reveal little in isolation, but together they can help construct a believable context. The answer is not to remove a legitimate public presence; it is to avoid unnecessary operational detail and train staff to expect targeted approaches.

Business email compromise remains a practical priority

One of the most important AI-related risks is not exotic malware but ordinary fraud carried out with better context. If an attacker gains access to a genuine mailbox, they may observe normal correspondence and identify an approaching payment, property transaction, acquisition, supplier settlement or other financially significant event. AI can assist in summarising the thread and understanding unfamiliar commercial language.

The resulting fraudulent instruction can appear within a genuine conversation and contain details known only to the parties. A request to change bank details should therefore never become credible merely because the surrounding email is authentic.

Independent callback, known contact details, segregation of duties and dual approval for significant transactions remain disproportionately valuable because they protect the process the criminal needs to exploit.

Voice and video are useful signals, not proof of identity

Synthetic speech, images and video can support impersonation. A telephone call or video appearance may add confidence without providing certainty. Organisations should retain procedural controls even where a supposed director, client or supplier appears personally to request an exception.

The resilient approach is procedural: authenticate unusual requests using trusted contact routes, keep dual controls for important payments and do not allow urgency or seniority to override the process. A familiar voice is easier to imitate than a well-designed approval system.

The technical threat is developing, but human operators still matter

AI coding systems can explain unfamiliar software, analyse errors and assist with programming. Those capabilities can also help skilled attackers with vulnerability research and tooling. The NCSC expects AI-assisted vulnerability research and exploit development to be among the most significant near-term developments, and warns that AI will further compress the time between disclosure of known vulnerabilities and exploitation. ^[1]

There are signs of more advanced use. In May 2026, Google Threat Intelligence Group reported the first case in which it had identified a threat actor using a zero-day exploit that it believed had been developed with AI assistance. ^[4] In July 2026, the NCSC and international partners reported that technical analysis indicated AI had played a role in developing a simple codebase used in a Russian state-supported phishing campaign targeting Zimbra users. ^[7]

These developments warrant attention without implying that autonomous AI is independently compromising organisations at will. The NCSC assesses fully automated end-to-end advanced cyber attacks as unlikely to 2027; skilled actors are expected to remain in the loop even as more individual stages become automated. ^[1]

Practical implication: patching speed matters. Known internet-facing vulnerabilities are likely to face a progressively shorter window before automated discovery and exploitation.

AI systems themselves create an additional attack surface

There is a second issue distinct from attackers using AI: organisations are giving AI systems access to their own information and tools. A chatbot with no access to company systems presents a different risk from an agent that can read email, search internal documents, call APIs, update records or take actions in other applications.

The more authority an AI agent has, the more consequential manipulation becomes. Prompt injection is one example. Untrusted material in an email, document or webpage can contain instructions intended to influence an AI system processing that content. The NCSC cautions that prompt injection should not be treated as though it can simply be eliminated in the same way as conventional SQL injection; the emphasis should be on reducing likelihood and limiting impact. ^[6]

For agentic AI, the NCSC recommends familiar security principles: least privilege, constrained scope, temporary credentials where possible, secure defaults, dependency management, monitoring, threat modelling and incident planning. ^[5]

Indicators that deserve contextual assessment

AI may improve presentation, but many underlying warning signs remain familiar:

- unexpected pressure for secrecy, urgency or an exception to normal procedure;
- a change to bank details or payment destination, especially late in a transaction;
- authentication notifications or password resets the user did not initiate;
- new email forwarding rules, unusual logins or unexplained session activity;
- requests to install software, share credentials or move a conversation to an unusual platform;
- a supplier or customer reporting messages that the organisation did not send;
- an AI agent making unexpected tool calls, accessing unusual data or attempting actions outside its normal workflow.

Where the evidence is often overstated

The objective should not be to turn staff into AI detectors. In many cases they will not be able to tell whether text, audio or code involved AI, and that distinction may have little relevance to the immediate response.

Nor does AI make every attacker highly capable. Public threat reporting repeatedly shows continued reliance on traditional infrastructure and human operators. The useful security assumption is simply that common attack tasks are getting cheaper and faster.

That makes established controls more valuable, not less. Strong authentication, rapid patching, independent payment verification, access control, monitoring and clear escalation procedures are difficult to defeat merely by producing more polished content.

A proportionate conclusion

AI is becoming a meaningful force multiplier in cybercrime and hostile cyber operations. It can accelerate reconnaissance, improve social engineering, assist coding and vulnerability research, process stolen information and increasingly automate individual stages of an intrusion. At the same time, the evidence does not support treating AI as a universal autonomous hacking capability.

For most businesses the proportionate response is practical: assume communications can be convincingly fabricated, reduce the value of public operational information, protect accounts, verify consequential instructions, patch exposed systems quickly and tightly govern any AI system that can access data or take actions.

The strongest defence against AI-assisted cybercrime is likely to remain less exotic than the technology itself: well-designed processes, properly configured systems and people who know when something warrants an independent check.

Real-world context: AI changed the credibility of impersonation

In 2024 Arup confirmed a fraud in which an employee in Hong Kong transferred HK\$200 million after a video conference using digitally generated representations of senior colleagues. The significance is not that every fraud now requires sophisticated AI, but that visual and voice familiarity can no longer be treated as independent proof that an instruction is genuine.

Public source: Arup deepfake fraud, publicly reported and confirmed in May 2024.

Sources and further reading

Sources were accessed and checked for this publication on 7 August 2026. Public threat reporting describes observed activity and assessment at the date of publication; capabilities and campaigns continue to evolve.

1. **National Cyber Security Centre.** *Impact of AI on cyber threat from now to 2027.* 7 May 2025. [Read source](#)
2. **OpenAI.** *Disrupting malicious uses of AI.* 25 February 2026. [Read source](#)
3. **Microsoft Threat Intelligence.** *AI as tradecraft: How threat actors operationalize AI.* 6 March 2026. [Read source](#)
4. **Google Threat Intelligence Group.** *GTIG AI Threat Tracker: Adversaries Leverage AI for Vulnerability Exploitation, Augmented Operations, and Initial Access.* 11 May 2026. [Read source](#)
5. **National Cyber Security Centre.** *Thinking carefully before adopting agentic AI.* 15 May 2026. [Read source](#)
6. **National Cyber Security Centre.** *Prompt injection is not SQL injection (it may be worse).* 8 December 2025. [Read source](#)
7. **National Cyber Security Centre.** *UK and partners expose Russian state-supported actors for new 'zero-click' phishing campaign.* 23 July 2026. [Read source](#)

Scope note

This publication provides general information and analysis. It is not legal, regulatory, financial, cyber-security or other professional advice. Specific incidents require appropriate technical, legal and operational advice based on the facts and jurisdiction.