VERIFY CARRATU
INTERNATIONAL LTD

RESEARCH & ANALYSIS

AI-ASSISTED CYBERCRIME / 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.

BY VCI / 7 AUGUST 2026 / 14 MINUTE READ

INTELLIGENCE, INVESTIGATION
AND DUE DILIGENCE

Independent analysis for decisions where the evidence, context and consequences matter.

VERIFY CARRATU INTERNATIONAL LTD

COMPANY NO. 17327582 / ICO REGISTRATION: ZC202007

CARRATU.CO.UK

LONDON / GENEVA

AI-Assisted Cybercrime: What Businesses Should Watch For

EVOLUTION	HUMAN IN THE LOOP	ZERO-DAY SIGNAL	AGENT ACCESS
AI is expected mainly to enhance existing attack methods, not replace them.	**Fully automated end-to-end advanced attacks remain unlikely in the near term.**	**Google reported a 2026 zero-day exploit believed to have been developed with AI assistance.**	**As AI agents gain tools and data access, least privilege and human approval matter more.**

The change is acceleration, not reinvention

Artificial intelligence is often discussed as though it has created an entirely new generation of cyber threats. The more useful view is less dramatic. Most successful attacks still depend on familiar weaknesses: compromised credentials, deceptive communications, unpatched software, excessive privileges, poorly configured systems and human error.

What AI changes is the amount of 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 enhancement of existing tactics rather than wholly new threat vectors.

Practical implication: do not focus on whether an attack was "made by AI". Ask whether existing controls still work when an attacker can research, write, translate and adapt faster.

Where AI can assist an attacker

Stage	How AI changes the task	Defensive emphasis
Reconnaissance	Combining public information about people, suppliers, technology and relationships.	Reduce unnecessary operational disclosure.
Social engineering	Drafting polished, tailored messages in multiple languages and many variants.	Judge the requested action, not the writing quality.
Coding & vulnerability research	Explaining unfamiliar software, debugging code and accelerating research.	Patch promptly, especially exposed services.
Post-compromise triage	Classifying files, summarising correspondence and identifying valuable data.	Limit access and monitor unusual downloads.

A realistic AI-assisted attack chain

AI is best understood as an assistant at several points in an otherwise conventional attack. The sequence below is illustrative rather than a technical recipe.

1	Public information	Legitimate public information is gathered about the organisation, its people, suppliers and technology.
2	AI-assisted profiling	AI helps sort the material, identify relationships and produce a coherent picture of who may trust whom.
3	Tailored contact	A message is written around a plausible invoice, document, supplier query or executive request.
4	Account or process compromise	The victim is persuaded to disclose access, approve an action or rely on a compromised route.
5	AI-assisted analysis	If access is obtained, AI can help triage correspondence, documents or unfamiliar systems.
6	Fraud, theft or further access	The objective remains conventional: money, data, persistence or another victim.

Phishing is becoming harder to recognise by appearance alone

Poor spelling and awkward grammar can still be warning signs, but they are no longer dependable. Generative AI can produce fluent correspondence, adapt tone and terminology, translate approaches and create many variants. The safer habit is to assess the requested action. Changes to bank details, credential requests, software installation, unexpected payments or requests to bypass a normal process warrant verification however professional the message appears.

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.

Business email compromise remains a practical priority

If an attacker gains access to a genuine mailbox, they may observe normal correspondence and identify an approaching payment or other financially significant event. AI can assist in summarising the thread and understanding unfamiliar commercial language. The fraudulent instruction can then appear inside a genuine conversation. Independent callback, known contact details, segregation of duties and dual approval remain disproportionately valuable controls.

Voice and video are useful signals, not proof of identity

Synthetic speech, images and video can support impersonation. A familiar voice or video appearance may increase confidence without proving identity. Retain procedural controls even where a supposed director, client or supplier appears personally to request an exception. 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 compress the time between disclosure of known vulnerabilities and exploitation.

In May 2026 Google Threat Intelligence Group reported a case in which it identified a threat actor using a zero-day exploit believed to have been developed with AI assistance. In July 2026 the NCSC and international partners reported that AI had played a role in developing a simple codebase used in a Russian state-supported phishing campaign targeting Zimbra users. These developments warrant attention without implying that autonomous AI is independently compromising organisations at will.

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

AI systems themselves create an additional attack surface

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. Prompt injection is one example: untrusted content may contain instructions intended to influence an AI system processing it. The appropriate response is familiar security engineering - least privilege, constrained scope, secure defaults, monitoring, human approval for consequential actions and incident planning.

What businesses should do now

Protect identities	Use strong MFA or passkeys for email, finance and administrative accounts.
Verify instructions	Check bank-detail changes, credential requests and unusual executive instructions independently.
Keep dual controls	Do not let urgency, hierarchy or convincing media bypass significant payment approvals.
Patch promptly	Prioritise internet-facing services, remote access, email platforms and edge devices.
Monitor accounts	Watch for unexpected logins, forwarding rules, unusual downloads and session changes.
Apply least privilege	Give employees, applications and AI agents only the access required for the task.
Review public exposure	Consider what websites, recruitment adverts and social posts reveal when combined.
Protect recovery	Maintain tested backups and ensure critical information can be restored.
Govern AI use	Define permitted data, integrations and actions that require human approval.
Prepare escalation	Make suspicious messages, prompts and payment requests easy to report quickly.

Warning signs worth taking seriously

•	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 prompts, password resets, forwarding rules or logins the user did not initiate;
•	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 or attempting actions outside its normal workflow.

What not to overreact to

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. The useful assumption is simply that common attack tasks are becoming cheaper and faster. Established controls therefore become more valuable, not less.

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 information and automate individual stages of an intrusion. The evidence does not support treating it as a universal autonomous hacking capability. For most businesses the proportionate response is practical: protect accounts, verify consequential instructions, patch exposed systems quickly and tightly govern AI systems that can access data or take actions.

Sources and reading

Sources were accessed and checked for this edition on 7 August 2026. Public threat reporting describes observed activity and assessment at the date of publication.

1. National Cyber Security Centre. Impact of AI on cyber threat from now to 2027. 7 May 2025. ncsc.gov.uk
2. OpenAI. Disrupting malicious uses of AI. 25 February 2026. openai.com
3. Microsoft Threat Intelligence. AI as tradecraft: How threat actors operationalize AI. 6 March 2026. microsoft.com
4. Google Threat Intelligence Group. AI Threat Tracker: vulnerability exploitation and initial access. 11 May 2026. cloud.google.com
5. National Cyber Security Centre. Thinking carefully before adopting agentic AI. 15 May 2026. ncsc.gov.uk
6. National Cyber Security Centre. Prompt injection is not SQL injection (it may be worse). 8 December 2025. ncsc.gov.uk
7. National Cyber Security Centre. UK and partners expose Russian state-supported actors for new 'zero-click' phishing campaign. 23 July 2026. ncsc.gov.uk

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.

About VCI

VCI provides investigation, intelligence and due diligence support to organisations, professional advisers and private clients. Work is scoped around the decision, the evidence available and the need for proportionate, clearly qualified reporting.