ChatGPT

The Coming AI Cybersecurity Crisis

Executive summary

Most companies are **not yet ready** for an AI-shaped cyber threat environment. The strongest current evidence does **not** suggest that AI has already replaced conventional intrusion tradecraft; rather, it shows that AI is making familiar attacks **cheaper, faster, more scalable and harder to detect**, especially social engineering, credential theft, data analysis, and vulnerability exploitation. At the same time, many organisations are deploying AI tools faster than they are governing them. In a 2025 survey cited by the World Economic Forum [1], **66%** of organisations expected AI to have the biggest impact on cybersecurity in the following year, but only **37%** said they had processes to assess the security of AI tools before deployment. The same research found that only **14%** were confident they had the cyber skills they need today. [2]

The immediate risk is not a Hollywood-style fully autonomous AI war. It is a more practical and more dangerous combination: **AI-enhanced phishing, vishing, deepfake impersonation, faster exploitation of known flaws, shadow-AI data leakage, and attacks on enterprise AI systems themselves** through prompt injection, data poisoning, model abuse, and supply-chain compromise. The NCSC [3] assessed in 2024 that AI would almost certainly increase the volume and impact of cyber attacks over the next two years, and in 2025 it warned that by 2027 AI-enabled tools would almost certainly accelerate exploitation of known vulnerabilities and raise risk to critical national infrastructure and its supply chains. [4]

Publicly documented incidents now demonstrate that the threat is not hypothetical. A finance worker in Hong Kong [5] was duped in early 2024 into transferring **HK$200 million** via a deepfake video conference scam that was later publicly identified as affecting Arup [6]. In 2025, the FBI [7] warned that malicious actors were using AI-generated voice to impersonate senior US officials. In late 2025, Anthropic [8] reported what it described as the first documented AI-orchestrated cyber-espionage campaign, targeting roughly 30 entities and validating a handful of successful intrusions. [9]

Readiness is uneven. **Finance** is the most mature large sector in governance and testing, largely because of regulatory pressure and mature resilience practices, but it remains highly exposed to fraud, third-party concentration risk, and DDoS. **Healthcare** and **manufacturing** are the least ready on a relative basis because of legacy environments, limited downtime tolerance, and weaker control coverage outside leading organisations. **Energy** has better strategic attention and more formal baselines, but its operational-technology exposure and long asset life create structural risk. **SMBs** are the least prepared overall because they lack specialist staff, formal response plans, and procurement leverage. These are analytical readiness judgements based on the best available public evidence; there is no single cross-sector global benchmark for “AI cyber readiness”. [10]

The good news is that the control agenda is unusually clear. The strongest evidence points to a relatively small set of high-value measures: **phishing-resistant authentication, rigorous AI inventory and access controls, external attack-surface reduction, third-party AI and software supply-chain governance, secure deployment of AI systems, deepfake-resistant financial workflows, faster patching, exercised incident response, and structured board oversight**. Existing frameworks from

NIST [11], ISO [12], MITRE [13], NCSC and CISA [14] are already sufficient to start; the gap is less a lack of frameworks than a lack of disciplined execution. [15]

Executive recommendations

- **Treat AI cyber risk as a core enterprise risk, not a tooling issue.** Put AI security and AI governance under the same executive risk structure, with explicit ownership for approved use, third-party review, and incident escalation. [16]
- **Prioritise identity and workflow controls first.** In the next twelve months, phishing-resistant MFA/passkeys, privileged-access hardening, callback verification for money movement, and device and session trust will usually reduce more risk than buying another "AI security platform". [17]
- **Inventory all AI use, including shadow AI.** Most firms still do not know which employees, teams, suppliers or applications are moving sensitive data into AI systems. That visibility gap is now one of the most important control failures. [18]
- **Build for shorter attacker cycle times.** Assume disclosed vulnerabilities will be weaponised faster, assume spear-phishing will be more personalized, and assume data stolen from one system will be quickly analysed and re-used elsewhere. [19]
- **Adopt threat-led testing for both classic IT/OT estate and AI components.** Red-team the human process, the AI application, the model supply chain, and the crisis-response process together. [20]

Threat landscape

The most important distinction is between **AI used by attackers** and **attacks on AI systems**. The first category is already material today: threat actors are using AI to generate convincing text, translate and localise lures, summarise stolen data, automate reconnaissance, debug scripts, write attack infrastructure, and support malware development. OpenAI [21] and Microsoft [22] reported in early 2024 that five state-affiliated groups had used LLM services for open-source research, translation, coding tasks and phishing support. OpenAI's 2025 reporting continued to find that most threat actors were still "bolting AI onto old playbooks" to move faster rather than gaining radically new capabilities. [23]

The second category is becoming more important as companies embed copilots, RAG pipelines, code agents and model APIs into production systems. NIST's adversarial machine learning taxonomy, the draft Cyber AI Profile, and NCSC's secure AI guidance all point to a common set of risks: prompt injection, indirect prompt injection, poisoning of training or retrieval data, model evasion, model extraction, leakage of prompts or sensitive data, and compromise of the broader AI supply chain, including models, datasets, plugins, APIs, tooling and infrastructure. [24]

A key analytical judgement from the evidence is that **AI disproportionately strengthens the attack stages where labour and time used to be the main constraints**: target research, lure creation, message variation, data triage, and adaptation. That matters because it shifts attacker economics. The NCSC's 2025 assessment that AI will likely shrink the time from vulnerability disclosure to exploitation is especially important for defenders, because most organisations are still failing to remediate exposure quickly enough. Verizon's 2025 DBIR found that only about **54%** of edge-device vulnerabilities were fully remediated during the year studied, with a median of **32 days** to complete remediation. [25]

Threat category	Typical attack vectors	Why AI changes the risk	Public evidence
AI-enhanced social engineering	Phishing, vishing, BEC, deepfake video calls, fake recruiters, impersonation of executives or officials	Better grammar, localisation, personalisation, voice cloning, rapid content variation, lower marginal cost per target	NCSC warns AI will increase phishing and ransomware impact; WEF recorded sharp growth in phishing/social engineering in 2024; IBM and law-enforcement reporting link deepfakes to fraud and impersonation. [26]
AI-assisted exploitation and intrusion support	Reconnaissance, exploit research, code generation, debugging, post-compromise scripting	Reduces skill threshold and accelerates attacker workflow; likely shrinks disclosure-to-exploit window	OpenAI/Microsoft early cases; NCSC 2025 forecast; Anthropic's 2025 campaign report; Google GTIG report. [27]
AI-enabled malware and evasion	Malware generation, obfuscation, dynamic script creation, adaptive C2 logic	Emerging move from "AI for productivity" to "AI in the malware loop" itself	Google GTIG reported the first "just-in-time" AI use in malware families PROMPTFLUX and PROMPTSTEAL. [28]
Attacks on enterprise AI applications	Prompt injection, indirect prompt injection, RAG poisoning, tool misuse, agent hijacking, data exfiltration	AI apps collapse application, data, identity and human-trust boundaries into one control plane	NIST AML taxonomy and Cyber AI Profile explicitly treat AI-specific attack surfaces as distinct from standard cyber risk. [29]
Shadow AI and data leakage	Employees using unsanctioned AI tools, weak authentication to GenAI apps, unmanaged browser or device usage	Sensitive data can leave the enterprise invisibly and be mixed with weak governance and poor logging	IBM found 97% of organisations reporting an AI-related security incident lacked proper AI access controls; Verizon observed routine employee access to GenAI systems and significant non-corporate-account usage. [30]
AI and software supply-chain abuse	Compromised models, poisoned datasets, insecure code assistants, third-party plugins, dependency attacks	AI expands the supplier graph and adds opaque dependencies that traditional vendor reviews miss	NIST AI 800-1 emphasises misuse risk across the AI supply chain; CISA has issued dedicated software supply-chain defence guidance. [31]

Threat category	Typical attack vectors	Why AI changes the risk	Public evidence
Public-sector impersonation and influence operations	Deepfake audio/video, bot orchestration, AI-generated content, credential harvest via impersonation	Faster content creation and more scalable networked deception	Anthropic documented multi-platform influence operations; FBI warned of AI-generated voice impersonation of senior officials. [32]

The current balance of evidence suggests that **social engineering remains the highest-probability enterprise risk in the near term**, while **AI-enabled exploitation and AI-specific application abuse are the fastest-rising technical risks**. The transition point to watch is the move from AI as an assistant to AI as an operational agent inside attack chains. Google's and Anthropic's late-2025 reports suggest that this transition has begun, even if it is not yet the norm. [33]

Documented incidents and case studies

The public record is still thin relative to the likely scale of real-world activity. Many victims do not disclose whether AI was used; even when they do, forensic proof is hard. The result is that the "incident" evidence base is biased toward law-enforcement statements, public company admissions, and the threat-intelligence reporting of major AI vendors and security firms. That is a significant data limitation, but it is no longer plausible to argue that AI-enabled cyber abuse is merely speculative. [34]

Date	Incident / milestone	Sector	What AI enabled	Reported impact
End-Jan 2024 / public in Feb-May 2024	Hong Kong deepfake video conference fraud later publicly identified as affecting Arup	Private sector	Deepfake impersonation of a CFO and colleagues in a fake video conference, following a phishing email	Hong Kong authorities said losses reached **HK$200 million**; Arup said operations and financial stability were not affected. [35]
Feb 2024	OpenAI and Microsoft disruption of five state-affiliated actors using LLM services	Public/ private targets	OSINT, translation, coding error finding, basic scripting, phishing support	Shows AI already embedded in state-linked cyber workflows, even if mostly as augmentation rather than breakthrough capability. [36]
Mar 2025	Anthropic reports recruitment-fraud enhancement and malware development by a low-skill actor	Mixed	Scam-content enhancement; malware creation support	Anthropic did not confirm successful deployment, but the case illustrates reduced skill barriers. [37]

Date	Incident / milestone	Sector	What AI enabled	Reported impact
May 2025	FBI warning on malicious messaging campaign impersonating senior US officials	Public sector	AI-generated voice plus trust-building messages to steal access and pivot to other targets	The FBI warned current and former federal and state officials and their contacts, highlighting credential and account-theft risk. [38]
Aug 2025	DARPA [39] AI Cyber Challenge finals	Defensive milestone	AI systems autonomously searched code and generated patches	Finalists found **18 real, non-synthetic vulnerabilities** and produced **11 patches**, showing how quickly autonomous cyber capability is improving on the defensive side too. [40]
Nov 2025	Anthropic's GTG-1002 report on an AI-orchestrated espionage campaign	Public/ private mixed	AI-assisted recon, vulnerability discovery, exploitation, lateral movement, credential harvesting, data analysis and exfiltration, allegedly with 80–90% of tactical operations delegated to AI	Anthropic says roughly **30 entities** were targeted and a **handful of intrusions** were validated as successful. [41]
Nov 2025	Google [42] Threat Intelligence Group reports first "just-in-time" AI-enabled malware	Cross-sector	Malware dynamically calling LLMs during execution to generate scripts or malicious functions	Important milestone because it suggests a shift from AI-assisted operators to AI-assisted malware behaviour. [28]

```
timeline
    title Selected AI-enabled cyber milestones
    2024-01 : NCSC warns AI will almost certainly increase the volume and
impact of cyber attacks
    2024-01 : Hong Kong police receive deepfake video-conference fraud report
    2024-02 : OpenAI and Microsoft disrupt five state-affiliated actors using
LLMs
    2025-05 : FBI warns of AI-generated voice impersonation of senior
officials
    2025-08 : DARPA AIxCC finalists discover real vulnerabilities and patches
    2025-11 : Anthropic reports AI-orchestrated cyber-espionage campaign
    2025-11 : Google reports first "just-in-time" AI-enabled malware
```

The timeline shows a clear progression: **warning**, then **fraud**, then **workflow augmentation**, then **operational abuse**, and finally **partial autonomy inside live intrusion chains**. The evidence is still incomplete, but the trend is consistent across government, frontier-model providers and mainstream threat-intelligence reporting. [43]

Corporate readiness by industry

The table below is an **analytical synthesis**, not a standardised league table. "Readiness" here means the likely ability of a typical organisation in each sector to **govern AI use, resist AI-enabled attack, detect compromise quickly, and recover without disproportionate business harm**. Public evidence allows stronger judgement on some sectors than others, and the data are skewed toward the US, UK and EU.

Industry	Relative readiness	Why this is the best current judgement	Main AI-cyber pressure points
Finance	**Medium-high**	Strongest regulatory and testing pressure. The Bank of England [44] continues to emphasise proactive detection, continuous monitoring and intelligence sharing, and the EU's DORA has applied since 17 January 2025. But finance is still heavily targeted for DDoS, fraud, social engineering and provider compromise. [45]	Deepfake fraud, AI-assisted scams, third-party concentration, customer impersonation, DDoS, model risk in core decisioning
Healthcare	**Low-medium**	Strategic attention is rising, but the control baseline remains weak and uneven. HHS's own inspector general found HIPAA audits too narrowly scoped to assess technical safeguards effectively; HHS also had to issue broad guidance around the Change Healthcare incident, which by 2025 affected about **192.7 million** individuals. Smaller facilities often have limited IT resources. [46]	Patient-safety disruption, legacy clinical systems, biomedical device exposure, business-associate risk, downtime complexity
Energy	**Medium**	Critical-infrastructure focus and sector baselines are improving. The IEA [47] reported that in 2024 a typical energy utility faced **more than 1,500 attempted cyberattacks per week**, triple the number four years earlier. DOE and sector bodies are establishing baselines for distribution systems and DERs, but OT/ICS exposure and long asset life mean structural residual risk remains high. [48]	AI-assisted exploitation of exposed OT, remote-access abuse, supply-chain compromise, long patch cycles, physical consequences

Industry	Relative readiness	Why this is the best current judgement	Main AI-cyber pressure points
Tech and SaaS	**Medium-high but uneven**	Large firms usually have the best cyber talent and the earliest access to AI security tooling, yet they also operate at the leading edge of exposure. ENISA found digital infrastructure and services to be the most targeted sector in its 2025 landscape, while WEF data show many organisations still deploy AI faster than they assess it. [49]	Software supply chain, AI app security, model-hosting abuse, prompt injection, shadow AI, customer concentration risk
Manufacturing	**Low-medium**	The sector remains structurally exposed because of OT, legacy environments, downtime intolerance and supplier sprawl. IBM reported manufacturing as the most attacked industry for multiple consecutive years; Verizon's manufacturing snapshot found more than **90%** of breached organisations in that dataset were SMBs, and ransomware pressure was acute. [50]	Ransomware, AI-assisted credential abuse, third-party secrets leakage, OT segmentation gaps, patching constraints
SMBs	**Low**	Small firms seldom have dedicated AI governance, deep vendor review, or mature IR. UK government survey data show cyber compromise remains common, while the US SBA says **88%** of small business owners felt vulnerable; the SBA also cited **$2.9 billion** in cybercrime cost to the small-business community in 2023. [51]	No in-house expertise, weak formal process, shadow AI, overreliance on MSPs, limited leverage over software suppliers

Two broad patterns stand out. First, the sectors with the most mature cyber programmes are **not** necessarily the most secure; they are often simply better at governance, testing and recovery. Finance illustrates this. Second, sectors with real-world dependencies on fragile or safety-critical operations—healthcare, energy and manufacturing—face the steepest downside if AI shortens attacker cycle times faster than they can shorten patching and containment times. [52]

Frameworks and governance

The framework landscape is not empty; it is crowded. The core challenge is integration. The most useful current stack starts with NIST's voluntary AI RMF and the general cyber baseline of CSF 2.0, then layers on AI-specific guidance for adversarial ML, GenAI, AI misuse, model supply chains and AI-enabled attacks. NIST's AI RMF 1.0 defines the four functions **Govern, Map, Measure and Manage**, while the GenAI Profile released in July 2024 adds technology-specific risk guidance. NIST's adversarial ML taxonomy provides a structured vocabulary for evasion, poisoning and privacy attacks, and the draft Cyber AI Profile reframes AI cyber work around three operational lenses: **Secure, Defend and Thwart**. [53]

NIST's January 2025 draft on misuse risk for dual-use foundation models is particularly important for boards and vendors because it recognises that **misuse risk has to be managed across the AI supply chain**, not just by model developers. That means cloud providers, downstream application builders, enterprises, and distribution platforms each have a role in reporting issues, sharing signals and controlling abuse. [54]

Outside NIST, the most relevant governance standards are **ISO/IEC 42001**, which specifies requirements for an AI management system, and **ISO/IEC 23894**, which gives AI-specific risk-management guidance. These are especially useful for firms that need a management-system approach rather than only technical controls. NCSC's secure AI system development guidance complements them by organising practice across secure design, secure development, secure deployment, and secure operation and maintenance. [55]

For threat modelling and operational detection, traditional MITRE ATT&CK remains useful for the **non-AI** parts of an intrusion chain, but it is incomplete for AI-specific tactics. That is why MITRE ATLAS matters: it is a living knowledge base of adversary tactics and techniques against AI systems. MITRE's SAFE-AI work explicitly connects AI-system elements to ATLAS threats, making it easier to adapt existing ATT&CK-based programmes rather than replacing them. [56]

CISA and NCSC already provide much of the non-negotiable baseline that enterprises need regardless of sector: secure-by-design procurement, secure-by-demand pressure on vendors, phishing-resistant MFA, software supply-chain defence, and guidance for severe cyber threat preparation and crisis response. The important analytical point is that **AI governance and cyber resilience should not be separated**. The best available frameworks all assume that AI risks become manageable only when they are embedded in mainstream cyber, risk, legal, data and procurement processes. [57]

Vendor tool landscape

The vendor market is moving fast, but the categories are already clear. There are **AI copilots for SOC workflows**, such as Microsoft [22] Security Copilot, Google SecOps' Gemini features, and CrowdStrike Charlotte AI; there are **AI-use visibility and control products**, such as Palo Alto's AI Access Security; and there is a growing set of **AI posture-management** and AI red-teaming products. These tools can improve triage speed, rule writing, threat-intel summarisation and sanctioned/unsanctioned AI visibility. They should, however, be treated as accelerants for an existing operating model, not substitutes for identity controls, telemetry, or human judgement. [58]

Gaps and mitigation priorities

The central technical gap is **visibility**. Many organisations cannot yet answer five basic questions with confidence: which AI systems are in use; what data flows into them; which external models, plugins or APIs they depend on; where logs are stored; and how a security team would investigate abuse of those systems. WEF, IBM and Verizon findings all point to the same pattern: AI adoption is outrunning policy, access control and monitoring. [59]

The next gap is **skills and organisational design**. Cyber teams, data teams, product teams, procurement, legal, compliance and HR often own different fragments of AI risk. WEF's AI-and-cybersecurity report explicitly calls for a cross-disciplinary AI risk function and an inventory of AI applications so that "shadow AI" and mission-critical supply-chain use can be seen and governed. This is a strong indication that the operating-model failure is at least as serious as the tooling failure. [60]

A third gap is **incident response discipline**. Many sectors remain better at recovering from ordinary IT failure than from cyber incidents that combine identity compromise, third-party service loss, media scrutiny, and operational downtime. Healthcare guidance from ASPR TRACIE stresses regular IR practice, collaborative planning with facility leadership, mission-critical inventories, segmentation, backups, and plans for prolonged downtime. Financial-sector testing by the Bank of England similarly stresses monitoring, credentials management, segmentation, remediation planning and intelligence sharing. [61]

A fourth gap is **supply-chain and procurement control**. In AI, the supplier graph is broader than most firms realise: models, foundation-model APIs, hosting layers, vector stores, data suppliers, code assistants, model gateways, observability tools, plugins, browser extensions, and open-source dependencies all matter. NIST AI 800-1 is unusually explicit on supply-chain roles, while CISA's secure-by-demand and software-supply-chain guidance argue that buyers must force better defaults from vendors rather than treating procurement as an afterthought. [62]

Recommended actions by stakeholder

Stakeholder	Highest-value actions in the next 90 days	Structural actions in the next 12–24 months
CISO	Build an enterprise AI inventory; classify sanctioned vs shadow AI; require phishing-resistant MFA for privileged and high-risk users; put deepfake-resistant verification into finance and service-desk workflows; add AI apps, model APIs and RAG data stores to logging, DLP and threat monitoring. [63]	Embed AI systems into vulnerability management, third-party risk, red teaming and crisis exercises; map ATT&CK and ATLAS use cases; implement AI-specific detection content and retrieval-data integrity checks. [64]
CEO	Set a firm-wide rule that no production AI deployment proceeds without security, legal and data review; tie AI rollout approvals to measurable security prerequisites; require monthly reporting on AI adoption, risk exceptions and high-risk suppliers. [16]	Fund identity modernisation, exposure management, supplier governance and crisis readiness before discretionary AI expansion; make AI cyber resilience part of operating, not innovation, budget. [65]
Board	Ask management for a single integrated view of cyber risk **from AI** and cyber risk **to AI systems**; demand evidence of tested response plans and supplier concentration controls. [66]	Move from annual oversight to scenario-based supervision: severe cyber event, AI-provider outage, deepfake-enabled fraud, and AI application data leak. Link management incentives to resilience metrics, not just AI-delivery milestones. [67]
CIO / CTO	Freeze unsanctioned plugins and agents that cannot meet logging, identity and data-handling requirements; ensure patching and secret management cover AI tooling and developer workflows. [68]	Standardise secure AI architecture patterns: gateway, logging, retrieval isolation, environment separation, prompt and response filtering, rollback, and dependency attestation. [69]

Actionable checklist

For most organisations, the highest-return execution sequence is:

- [ ] **Create a living AI asset register** covering vendors, models, plugins, agents, data stores, and business owners. [70]
- [ ] **Require phishing-resistant MFA/passkeys** for privileged users, admins, finance approvers, help-desk staff and remote-access workflows. [71]
- [ ] **Introduce deepfake-resistant verification** for payment changes, secrets resets, HR/payroll requests and VIP access exceptions. [72]
- [ ] **Put all AI traffic behind approved gateways or brokers** with logging, DLP, identity binding and policy enforcement. [73]
- [ ] **Extend vulnerability and configuration management to AI-adjacent infrastructure** and assume faster exploit windows. [25]
- [ ] **Review third-party contracts** for model hosting, incident notification, logging retention, training-data use, and abuse-reporting obligations. [74]
- [ ] **Red-team both the human process and the AI application**: prompt injection, poisoned retrieval content, tool misuse, approval-chain bypass and media response. [20]
- [ ] **Exercise severe-cyber and prolonged-downtime scenarios** with business units, comms, legal, procurement and executive leadership. [75]
- [ ] **Join or deepen threat-intelligence sharing** relevant to sector and geography. [76]
- [ ] **Measure what matters**: AI inventory coverage, sanctioned-vs-shadow use, privileged MFA coverage, time to patch exposed systems, time to detect, and time to isolate a compromised AI-connected workflow. [77]

Regulation, investment and outlook

The regulatory landscape is tightening, but unevenly. The most consequential global regulatory development is the European Union [78] AI Act: it entered into force on **1 August 2024**, prohibited-practice and AI-literacy provisions started applying on **2 February 2025**, GPAI obligations on **2 August 2025**, and the majority of rules begin applying on **2 August 2026**; rules for certain high-risk AI systems embedded in regulated products extend to **2 August 2027**. Alongside it, NIS2 now covers **18 critical sectors**, DORA has applied to financial entities since **17 January 2025**, and the Cyber Resilience Act entered into force in December 2024, with vulnerability-reporting obligations starting on **11 September 2026** and full application by **11 December 2027**. [79]

Outside the EU, the picture is more fragmented but still moving. In the US, the SEC [80] cyber-disclosure rules require public companies to disclose material cyber incidents quickly and to describe cyber risk management and board oversight. The FTC [81] has begun enforcing its impersonation rule and has explicitly linked new protections to AI-generated impersonation harms. In the federal supply chain, the DoD's CMMC programme is pushing stronger baseline controls across defence contractors. The likely practical effect is that boards and executives will face rising expectations not only to secure AI, but also to show due diligence on vendor choice, access control, logging, incident disclosure and governance. [82]

The investment case for executives is relatively straightforward. IBM's 2025 breach research puts the global average breach cost at about **$4.4 million**, reports that **97%** of organisations that had an AI-related security incident lacked proper AI access controls, and finds that extensive use of security AI was associated with about **$1.9 million** in breach-cost savings compared with organisations that did not use it extensively. That does **not** mean "buy more AI". It means the best returns seem to come when AI is

paired with **identity controls, data governance, security operations discipline and faster containment**, not when it is deployed in governance vacuum. [83]

Priority investment areas

In priority order, the available evidence suggests that executive cyber investment over the next three years should concentrate on **identity**, **visibility**, **response**, and **supplier assurance**. Identity means phishing-resistant MFA, device trust, and privileged-access hardening. Visibility means AI inventory, gatewaying, telemetry and attack-surface management. Response means tested playbooks, backups, segmentation and crisis rehearsal. Supplier assurance means secure procurement, software supply-chain checks, contractual rights to logs and notification, and concentration-risk review for critical providers. These areas appear repeatedly across CISA, NCSC, WEF, Bank of England and NIST materials. [84]

Outlook scenarios

```
flowchart LR
    A[Cheaper and better offensive AI] --> B[Higher attack volume and better
targeting]
    B --> C[More phishing, vishing, deepfakes and credential theft]
    B --> D[Faster recon and vulnerability exploitation]
    C --> E[Initial access]
    D --> E
    E --> F[Operational disruption and extortion]
    E --> G[Data theft and AI-system compromise]
    H[Shadow AI and weak supplier governance] --> E
    I[Strong identity, logging, segmentation and exercises] --> J[Faster
containment and lower losses]
    F --> K[Financial, legal and regulatory impact]
    G --> K
```

Baseline scenario. AI remains primarily an amplifier of familiar threats. Most organisations see more convincing social engineering, faster exploitation of public-facing flaws, and more complicated insider/data-governance problems. Breach frequency rises modestly; the bigger effect is a higher tempo of attempted compromise and lower attacker operating cost. This is the scenario most strongly supported by NCSC, WEF and OpenAI's public reporting. [85]

Adverse scenario. Agentic AI systems move from assisting operators to coordinating bigger segments of the intrusion lifecycle, including vulnerability discovery, tool use, malware adaptation, and post-compromise operations. High-impact incidents become more clustered in sectors with OT, brittle supplier networks or weak identity controls. Anthropic's GTG-1002 report and Google's "just-in-time" malware findings are early warning signs for this path. [86]

Resilience scenario. Defenders adopt the best of offensive automation for defence. That means wider use of autonomous code analysis, attack-surface reduction, AI-supported detection engineering, exposure discovery and prioritised remediation. DARPA's AI Cyber Challenge shows that AI-driven vulnerability discovery and patch generation are already operationally meaningful on the defensive side. If paired with secure-by-design procurement, AI gateways, and tested crisis management, this scenario can offset much of the attacker productivity gain. [87]

Open questions and limitations

There are three major uncertainties. First, **under-reporting**: many real incidents probably go unlabelled as AI-enabled, so public case counts understate prevalence. Second, **telemetry bias**: some of the clearest evidence comes from AI providers and security vendors, which have unique visibility but also incomplete visibility and commercial incentives. Third, **rapid capability change**: several of the most relevant NIST and AI-security artefacts are still evolving, and frontier-model cyber capability can move faster than conventional policy cycles. Those uncertainties argue for humility in forecasting exact incident rates, but they do **not** weaken the core conclusion that the risk curve is bending upward and that execution discipline now matters more than waiting for clearer proof. [88]

In practical terms, the “coming AI cybersecurity crisis” is already here in its early phase. The organisations most likely to cope well are not necessarily those with the most AI, but those that can **govern it, constrain it, observe it, and keep operating when it fails or is abused**. [89]

---

[1] [10] [20] [21] [45] [52] [76] https://www.bankofengland.co.uk/financial-stability/operational-resilience-of-the-financial-sector/2025-cbest-thematic
https://www.bankofengland.co.uk/financial-stability/operational-resilience-of-the-financial-sector/2025-cbest-thematic

[2] [59] [66] [81] https://www.weforum.org/publications/global-cybersecurity-outlook-2025/digest/
https://www.weforum.org/publications/global-cybersecurity-outlook-2025/digest/

[3] [6] [24] [29] https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2025.pdf
https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2025.pdf

[4] [26] [43] https://www.ncsc.gov.uk/sites/default/files/pdfs/publication/impact-of-ai-on-cyber-threat.pdf
https://www.ncsc.gov.uk/sites/default/files/pdfs/publication/impact-of-ai-on-cyber-threat.pdf

[5] [11] [16] [18] [34] [60] [70] [89] https://reports.weforum.org/docs/WEF_Artificial_Intelligence_and_Cybersecurity_Balancing_Risks_and_Rewards_2025.pdf
https://reports.weforum.org/docs/WEF_Artificial_Intelligence_and_Cybersecurity_Balancing_Risks_and_Rewards_2025.pdf

[7] [14] [30] [39] [63] [65] [73] [77] [83] https://www.ibm.com/reports/data-breach
https://www.ibm.com/reports/data-breach

[8] [17] [71] https://www.cisa.gov/sites/default/files/publications/fact-sheet-implementing-phishing-resistant-mfa-508c.pdf
https://www.cisa.gov/sites/default/files/publications/fact-sheet-implementing-phishing-resistant-mfa-508c.pdf

[9] [35] [72] [78] https://www.info.gov.hk/gia/general/202406/26/P2024062600192.htm
https://www.info.gov.hk/gia/general/202406/26/P2024062600192.htm

[12] [46] https://oig.hhs.gov/reports/all/2024/the-office-for-civil-rights-should-enhance-its-hipaa-audit-program-to-enforce-hipaa-requirements-and-improve-the-protection-of-electronic-protected-health-information/
https://oig.hhs.gov/reports/all/2024/the-office-for-civil-rights-should-enhance-its-hipaa-audit-program-to-enforce-hipaa-requirements-and-improve-the-protection-of-electronic-protected-health-information/

[13] [31] [54] [62] [74] https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd2.pdf
https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd2.pdf

[15] [53] https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf
https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

19 25 85 https://www.ncsc.gov.uk/sites/default/files/pdfs/publication/impact-ai-cyber-threat-now-2027.pdf
https://www.ncsc.gov.uk/sites/default/files/pdfs/publication/impact-ai-cyber-threat-now-2027.pdf

22 28 33 https://services.google.com/fh/files/misc/advances-in-threat-actor-usage-of-ai-tools-en.pdf
https://services.google.com/fh/files/misc/advances-in-threat-actor-usage-of-ai-tools-en.pdf

23 27 36 https://openai.com/index/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors/
https://openai.com/index/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors/

32 37 https://www.anthropic.com/news/detecting-and-countering-malicious-uses-of-claude-march-2025
https://www.anthropic.com/news/detecting-and-countering-malicious-uses-of-claude-march-2025

38 https://www.fbi.gov/investigate/cyber/alerts/2025/senior-us-officials-impersonated-in-malicious-messaging-campaign
https://www.fbi.gov/investigate/cyber/alerts/2025/senior-us-officials-impersonated-in-malicious-messaging-campaign

40 87 https://www.darpa.mil/news/2025/aixcc-results
https://www.darpa.mil/news/2025/aixcc-results

41 42 86 https://assets.anthropic.com/m/ec212e6566a0d47/original/Disrupting-the-first-reported-AI-orchestrated-cyber-espionage-campaign.pdf
https://assets.anthropic.com/m/ec212e6566a0d47/original/Disrupting-the-first-reported-AI-orchestrated-cyber-espionage-campaign.pdf

44 48 https://www.iea.org/commentaries/are-governments-better-positioned-to-respond-to-energy-security-risks-today-than-in-the-past
https://www.iea.org/commentaries/are-governments-better-positioned-to-respond-to-energy-security-risks-today-than-in-the-past

47 84 https://fidoalliance.org/cisa-secure-by-demand-guide-phishing-resistant-authentication-passkeys-by-default/
https://fidoalliance.org/cisa-secure-by-demand-guide-phishing-resistant-authentication-passkeys-by-default/

49 https://www.enisa.europa.eu/sites/default/files/2026-01/ENISA%20Threat%20Landscape%202025_v1.2.pdf
https://www.enisa.europa.eu/sites/default/files/2026-01/ENISA%20Threat%20Landscape%202025_v1.2.pdf

50 https://newsroom.ibm.com/2025-04-17-2025-ibm-x-force-threat-index-large-scale-credential-theft-escalates%2C-threat-actors-pivot-to-stealthier-tactics
https://newsroom.ibm.com/2025-04-17-2025-ibm-x-force-threat-index-large-scale-credential-theft-escalates%2C-threat-actors-pivot-to-stealthier-tactics

51 https://www.gov.uk/government/statistics/cyber-security-breaches-survey-20252026/cyber-security-breaches-survey-20252026
https://www.gov.uk/government/statistics/cyber-security-breaches-survey-20252026/cyber-security-breaches-survey-20252026

55 80 https://www.iso.org/standard/42001
https://www.iso.org/standard/42001

56 64 https://atlas.mitre.org/
https://atlas.mitre.org/

57 https://www.cisa.gov/securebydesign
https://www.cisa.gov/securebydesign

58 https://www.microsoft.com/en-in/security/business/ai-machine-learning/microsoft-security-copilot
https://www.microsoft.com/en-in/security/business/ai-machine-learning/microsoft-security-copilot

61 https://files.asprtracie.hhs.gov/documents/aspr-tracie-healthcare-system-cybersercurity-readiness-response.pdf
https://files.asprtracie.hhs.gov/documents/aspr-tracie-healthcare-system-cybersercurity-readiness-response.pdf

67 75 https://www.ncsc.gov.uk/sites/default/files/2026-04/Preparing-for-severe-cyber-threat-Why-leaders-must-act-now.pdf
https://www.ncsc.gov.uk/sites/default/files/2026-04/Preparing-for-severe-cyber-threat-Why-leaders-must-act-now.pdf

68 https://www.verizon.com/business/resources/infographics/2025-dbir-manufacturing-snapshot.pdf
https://www.verizon.com/business/resources/infographics/2025-dbir-manufacturing-snapshot.pdf

69 https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development
https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development

79 https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

82 https://www.sec.gov/resources-small-businesses/small-business-compliance-guides/cybersecurity-risk-management-strategy-governance-incident-disclosure
https://www.sec.gov/resources-small-businesses/small-business-compliance-guides/cybersecurity-risk-management-strategy-governance-incident-disclosure

88 https://openai.com/global-affairs/disrupting-malicious-uses-of-ai-october-2025/
https://openai.com/global-affairs/disrupting-malicious-uses-of-ai-october-2025/