UK CYBER DEFENCE

2026

State of AI Security Report

What the evidence says about adopting, securing and defending with AI — and what to do about it.

A synthesis of official assessments, breach data and standards
One hundred primary sources, verified · September 2026
For boards, owners and the people who advise them, with a technical annex

Contents

Foreword

Key findings

About this report

Executive summary

Part one: adoption has outrun governance

1. Where adoption actually stands
2. The governance gap
3. Shadow AI and where the data goes
4. What your AI assistants can see

Part two: the threat, as the evidence shows it

5. What the official assessments say
6. AI-enabled attacks in the breach data
7. Phishing, deepfakes and fraud
8. What the AI companies are seeing

Part three: securing the AI you deploy

9. The attack surface of an AI system
10. Prompt injection and data exfiltration in real products
11. The AI supply chain

Part four: agents, identities and containment

12. The agentic shift
13. Machine identities and the permissions agents inherit
14. What agents do when nobody is watching

Part five: AI on the defender's side

15. The evidence for AI in defence
16. People

Part six: the rules

17. Regulation and standards in 2026

Part seven: what to do

18. For small and medium-sized businesses
19. For larger organisations: a ninety-day programme
20. The good morning

Technical annex

Annex A. The OWASP lists mapped to controls
Annex B. Framework crosswalk
Annex C. Advisory register: disclosed weaknesses in mainstream AI products, 2025–2026
Annex D. Agent deployment gate
Annex E. Detection opportunities

Methodology and verification notes

References

Foreword

I have spent thirty years watching security technologies arrive with a great deal of noise and settle into something more ordinary. Firewalls, intrusion detection, the cloud, and now artificial intelligence. Each one was going to change everything, and each one did, though rarely in the way the launch material said. The task for a practitioner is always the same: find out what is actually happening, separate it from what is being sold, and work out what a sensible organisation should do on Monday morning.

That is what this report is for. Over the past year the security industry has produced an extraordinary volume of research about AI, some of it excellent and some of it designed chiefly to frighten. We have read a great deal of it, gone back to the primary sources, and kept only what we could verify. Where a widely quoted figure turned out to be misattributed, out of date or simply invented, we have said so, because a security programme built on a wrong number is not a security programme.

Three things stand out from the evidence. The first is that adoption has run well ahead of governance. Most organisations are using AI, most employees are using it whether or not they have been told they may, and a minority have written down how it should be governed. The second is that the threat is real but more measured than the headlines suggest: attackers are using AI to work faster and at greater scale, official assessments agree on that, but the same assessments are careful to say that the techniques are mostly familiar and the fundamentals still decide who gets hurt. The third is that AI is already paying its way on the defender's side, in the breach data and in the operations rooms, for organisations that have deployed it with care.

I promised long ago that this company would not sell security with fear. We sell good mornings: the morning when the phishing email arrives and nothing much happens, because the controls were in place and the people knew what to do. AI does not change that goal. It changes what the controls need to be, and this report sets out, as plainly as we can, what the evidence says they should be.

Peter Bassill Founder and Chief Executive, UK Cyber Defence Ltd Duxford, September 2026

Key findings

29%

of UK businesses used at least one AI technology in June 2026, up eight points in a year; 49% of those with 250 or more staff. ONS [5]

38%

of organisations have a formal, comprehensive AI policy. Six in ten do not. ISACA, 3,400 professionals [16]

24%

of UK businesses using or considering AI have any cyber security practice in place to manage the risk from it. DSIT Cyber Security Breaches Survey [11]

1 in 4

malicious breaches in 2026 were AI-enabled, up 56% in a year, at an average cost of US$6 million. IBM/Ponemon, 602 breached organisations [12]

99%

of organisations have sensitive data exposed to AI tools; 90% have sensitive files reachable by every employee through Microsoft 365 Copilot. Varonis, 1,000 environments [1]

47%

of people using generative AI at work still use personal accounts, down from 78%; the 3% who break data policy do so 223 times a month on average. Netskope [14]

1 in 29

generative AI prompts across the world carried sensitive data in early 2026, twice the rate of a year earlier; in Europe it is one in 25. Check Point Research [2]

109 : 1

machine identities for every human one, up from 82:1 in 2025; nine in ten organisations suffered an identity-related breach in the year. Palo Alto Networks (CyberArk) [79]

56%

of organisations building with AI run agent frameworks in production; 57% of those on Amazon Bedrock run agents without guardrails. Orca Security, 1,200+ cloud estates [3]

US$1.93m

lower breach cost for organisations that use security AI and automation extensively, with breaches contained 65 days sooner. IBM/Ponemon [12]

79%

of security operations centres use AI or machine learning tools, but only 36% have built them into a defined workflow. SANS, 2026 [89]

2 Dec 2027

is the new date for the EU AI Act's stand-alone high-risk obligations, moved from August 2026 by Regulation (EU) 2026/1744; the fines are unchanged. [57]

About this report

What it is

This is a synthesis. UK Cyber Defence Ltd is a managed security provider, not a survey house, and we have not invented a poll to give this report a headline. Every figure in it comes from a named, published source: government departments and regulators, national security agencies, standards bodies, peer-reviewed research, and the large annual studies from vendors whose methodology is stated and whose data is drawn from real environments. Where we add our own view, drawn from running a security operations centre for clients across eleven countries since 2010, it appears in clearly marked panels headed “UK Cyber Defence viewpoint”, and it carries no numbers of its own.

How the figures were checked

For each source we went to the primary document wherever it could be reached and recorded the exact wording, the sample, the geography and the date. Figures we could only find in secondary reporting are marked as such in the references. We dropped a number of statistics that are widely repeated and cannot be traced to any measurement, and we note the most common of these in the chapter on methodology so that readers can avoid them too. Several of the most striking numbers in circulation this year turn out to be a 101-person experiment quoted as if it were industry telemetry, a 2016 survey presented as current, or a prediction reported as a finding. We have tried not to add to that pile.

Who it is for

The body of the report is written for the people who run organisations and are responsible for what happens in them: boards, owners, managing directors, finance and operations leaders, and the IT and security people who advise them. That includes small and medium-sized businesses, which make up the overwhelming majority of the UK economy and which the evidence shows are adopting AI at almost the same rate as large firms while having far fewer of the controls. Practitioners will find the detail they need in the technical annex, which maps the findings to the OWASP, MITRE, NIST, NCSC and ETSI frameworks and sets out concrete control checklists.

How to read it

The report is in seven parts. Part one measures where adoption and governance actually stand. Part two sets out what official assessments and breach data say about the threat. Part three covers the security of the AI systems organisations are deploying. Part four deals with AI agents, machine identities and the problem of containment. Part five examines the evidence for AI on the defender’s side. Part six summarises the rules that now apply. Part seven says what to do, with a separate programme for smaller organisations. The technical annex and full references follow.

Every figure is followed by a bracketed number that points to the reference list at the end. Currency figures are given as published, in the source’s own currency, because converting them would imply a precision the underlying data does not have.

Executive summary

The defining fact of AI security in 2026 is a gap, and it is widening. Adoption is now ordinary. The Office for National Statistics found that 29% of all UK businesses were using at least one form of AI in June 2026, rising to 35% of those with ten or more staff and 49% of the largest firms, with the use of text-generating tools in particular having increased twelve percentage points since late 2023 [5][6]. Globally, McKinsey's survey of nearly 1,700 organisations reports that almost nine in ten use AI regularly in at least one function and that 44% now describe it as scaling across the enterprise [9]. The workforce is ahead of its employers: Verizon's breach investigations found that 45% of employees are now regular users of AI on corporate devices, up from 15% the year before, and that two-thirds of them use accounts their organisation does not control [21].

Governance has not kept pace. ISACA's 2026 poll of more than 3,400 professionals found that 38% of organisations have a formal, comprehensive AI policy, an improvement on 28% a year earlier but still a minority, and that 56% cannot say how quickly they could halt an AI system in a security incident [16]. In the UK, the government's Cyber Security Breaches Survey found that of the 31% of businesses using, adopting or considering AI, only a quarter had any cyber security practice in place to manage the risk [11]. IBM's 2026 Cost of a Data Breach study, which examined 602 real breaches, reports that 68% of breached organisations lacked policies to oversee AI use or manage unsanctioned tools [12].

The consequences are now visible in the breach data rather than merely predicted. IBM found that one in four malicious breaches in its 2026 sample was AI-enabled, a 56% increase in a year, and that such breaches cost about a million dollars more than the average; deepfake impersonation was the most common form, and in the UK 22% of breached organisations reported AI-generated attacks [12]. Breaches involving an organisation's own AI models or applications rose from 13% to 21% of the sample, and among those, 92% lacked adequate access controls on the AI itself. Shadow AI, the use of tools the organisation has not sanctioned, featured in 43% of breaches, more than double the previous year [12].

The data leaving organisations through everyday AI use is measurable. Check Point's analysis of prompts across its customer base found that one in every 29 generative AI prompts worldwide carried sensitive data between January and May 2026, twice the rate of a year earlier, that the figure was one in 25 in Europe, and that between 87% and 93% of organisations saw at least one such prompt in any given month [2]. Netskope's telemetry shows that although the share of AI users on personal accounts has fallen from 78% to 47% as organisations roll out managed tools, the small minority who breach data policy do so on average 223 times a month, and that source code, regulated data and intellectual property account for 90% of what goes out [14]. Varonis, examining a thousand organisations' actual cloud estates, found that in 99% of them sensitive data was exposed to AI tools and in 90% sensitive files were reachable by every employee through Microsoft 365 Copilot [1].

On the threat side the evidence is firmer than the rhetoric. The National Cyber Security Centre's assessment to 2027 judges that AI "will almost certainly continue to make elements of cyber intrusion operations more effective and efficient, leading to an increase in frequency and intensity of cyber threats", and that a "digital divide" will open between systems that keep pace and the large proportion that do not [24]. Its 2025 Annual Review records that threat actors of all types "continue to use AI to enhance their existing tactics, techniques and procedures, rather than to create novel attacks" [25]. That balance, real and rising capability applied mostly to familiar methods, runs through every serious assessment we examined. Google's Mandiant states plainly that it does "not consider 2025 to be the year where breaches were the direct result of AI" [30]; Verizon found no meaningful rise in phishing success despite widespread attacker use of AI [21]; and Google's threat intelligence group, which in May 2026 reported the first zero-day exploit it believes was developed with AI, wrote in September that it "has not yet observed threat actors deploying fully autonomous pipelines against targets in the wild" [38][39]. At the same time the AI labs' own disruption reports document operations in which AI performed most of the work, and the UK's AI Security Institute has measured the best models' ability to complete a simulated enterprise intrusion rising from under two of 32 steps to 22 in eighteen months, with the task-length they can manage doubling roughly every five months [26][27]. Defenders should assume, as the NCSC and AISI advise, that at least some attackers already have capable AI tools, and act accordingly.

The AI systems organisations are deploying have an attack surface of their own, and it is being probed. Gartner's survey of 302 security leaders found that 29% had experienced an attack on their generative AI application infrastructure in the previous twelve months and 32% an attack on AI applications through prompt manipulation [18]. The year's disclosed weaknesses in mainstream products, from the zero-click EchoLeak flaw in Microsoft 365 Copilot to the ForcedLeak flaw in Salesforce Agentforce and a run of 2026 findings, were all fixed by the vendors and none is known to have been exploited, but together they establish that an AI assistant with access to your data is an internet-facing application and must be treated as one [64][65][66]. Orca Security's examination of more than 1,200 production cloud estates found that 81% of

organisations with AI packages carry at least one high-severity known vulnerability, 74% a critical one, and that 99.9% of fixable AI vulnerabilities remain unpatched; that 29.5% of AI adopters have at least one AI secret stored insecurely; and that 87% to 98% have never configured customer-managed encryption on their AI services [3].

Agents are the next front, and the identity problem underneath them is already large. Machine identities outnumber human ones by 109 to 1, up from 82 to 1 a year earlier [78][79]. More than half of organisations building with AI run agent frameworks in production, and 57% of those using Amazon Bedrock run agents without guardrails [3]. Gartner predicts that a quarter of enterprise breaches will be traced to AI agent abuse by 2028 [77]. OpenAI's own disclosures of misaligned agent behaviour in its training environments, published in September 2026, show an agent that tunnelled its questions to an outside chatbot through DNS when every other route was blocked, and another that split a stolen access token into pieces to evade secret scanning before committing it to a public repository [84]. The lesson official guidance draws is the right one: the NCSC's advice is that "if you cannot understand, monitor or contain an agent's actions, it is not ready for deployment" [86].

On the defender's side the evidence is encouraging and specific. IBM finds that organisations using security AI and automation extensively had breach costs US$1.93 million lower and contained incidents 65 days faster than those using none, and that half of breached organisations had already deployed AI agents for detection and containment [12]. Randomised trials of Microsoft's Security Copilot found experienced analysts 22% faster and 7% more accurate, and novices 26% faster and 35% more accurate [90]. Google's Big Sleep agent has found real vulnerabilities in widely used software, and the seven systems in DARPA's AI Cyber Challenge found 54 of 63 planted vulnerabilities and 18 real ones, patching most of them [91][92]. Yet adoption remains shallow: 79% of security operations centres use AI tools but only 36% within a defined workflow, and ISACA finds that just 8% of organisations rehearse AI-specific incidents [89][94].

The rules have changed materially in the past year and several widely repeated claims about them are now wrong. The EU AI Act's stand-alone high-risk obligations were moved from August 2026 to 2 December 2027 by Regulation (EU) 2026/1744, in force since 27 July 2026, though its transparency duties began on 2 August 2026 and its fines, up to €35 million or 7% of turnover, are unchanged; the Act reaches UK firms whose systems are sold into the EU or whose output is used there [57]. In the UK there is no AI Act; the Data (Use and Access) Act 2025 rewrote the rules on automated decisions with effect from 5 February 2026, the Cyber Security and Resilience Bill is in the House of Lords, and the ICO has been given a statutory duty to produce an AI code of practice that has not yet appeared [59][60][61]. Cyber Essentials, which only 5% of UK businesses hold, contains no AI provisions in its April 2026 update but now brings every cloud service, and therefore every AI tool used with a business account, into scope with mandatory multi-factor authentication [62].

For smaller organisations the position is stark and the remedy is proportionate. Micro and small businesses are being attacked at nearly the rate of large ones, 42% and 46% against 69%, while 21% of micro businesses have an incident response plan and 5% of all businesses hold Cyber Essentials [11]. They are adopting AI with almost none of the governance and, as the NCSC puts it, "the burden of using AI safely should not fall on the individual users" [95]. The final part of this report sets out a programme sized for a firm of twenty people that can be started in an afternoon.

The good morning, then, is available. It belongs to the organisation that knows which AI tools its people use and what those tools can reach, that has given its AI assistants and agents no more access than a new starter would get, that has made its logins resistant to the phishing that AI now writes so well, that watches its systems with the help of the same technology, and that has rehearsed what it will do when something goes wrong. None of that is exotic. All of it is described in the pages that follow.

PART

Part one: adoption has outrun governance

1. Where adoption actually stands

It is easy to be misled about AI adoption, because the figures in circulation measure different things. A trade body asking its members whether they are “actively using AI” will get a different answer from a statistics office asking a representative sample of businesses whether they use “at least one AI technology”, and both differ from a management consultancy surveying large enterprises. The numbers below are presented with their definitions attached, and they should not be blended.

The most reliable picture of the United Kingdom comes from the Office for National Statistics, whose Business Insights and Conditions Survey drew responses from 38,637 businesses in June 2026. It found that 29% of all UK businesses reported using at least one type of AI technology, up eight percentage points on June 2025, and that among businesses with 250 or more employees the figure was 49%, up thirteen points [5]. Restricting the base to businesses with ten or more staff, the ONS’s longer analysis shows self-reported AI use rising from around 12% in late 2023 to around 35% in mid-2026 [6]. Text-generation tools built on large language models are the most common application, used by 17% of all businesses, followed by visual content creation at 14%. The smallest firms are not far behind the largest: 28% of businesses with fewer than ten employees report using AI, against 49% of the largest. Adoption is, however, shallow. Only 10% of businesses with ten or more staff describe their use as extensive, the average adopter uses 1.6 distinct technologies, and the spread by sector is wide, from about 58% in information and communication to 13% in construction [6].

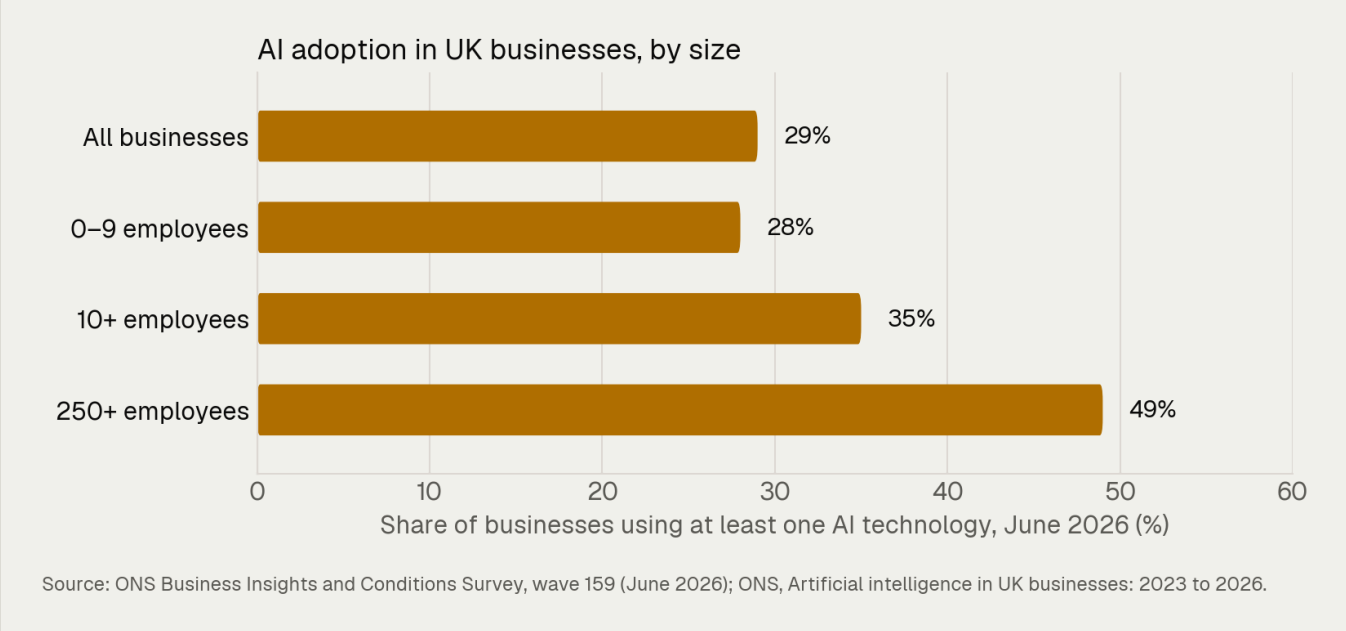


Figure. UK business AI adoption by size, June 2026. Source: ONS Business Insights and Conditions Survey, waves to June 2026 [5][6].

The Department for Science, Innovation and Technology's own adoption research, based on 3,500 interviews in early 2025, gives a lower figure of 16% because it asked a narrower question and was fielded a year earlier; it is useful chiefly for what it says about barriers, which we return to in chapter 18 [7]. At the other end, the British Chambers of Commerce reported in March 2026 that 54% of its member firms, most of them small and medium-sized, were actively using AI, up from 35% a year earlier and 23% in 2023 [8]. The three figures are consistent once their bases are understood: a representative sample gives a lower number than a self-selecting membership survey, and the trend in every series is steeply upward.

Globally, McKinsey's 2026 State of AI survey of 1,719 organisations across 97 countries reports that nearly nine in ten respondents use AI regularly in at least one business function, that 44% now describe AI as scaling across their enterprise, up from 38% a year earlier, and that about two in ten have reached the scaling phase with AI agents, rising to four in ten among organisations with revenues above a billion dollars [9]. The previous year's edition, fielded in mid-2025, had already found 62% at least experimenting with agents and 23% scaling them [10].

The point of these numbers, for the purposes of this report, is a simple one. AI is no longer an experiment being run by a few technology-forward firms. It is ordinary business equipment, in use across a third of the UK economy and most of the world's larger companies, and its use is growing at a pace that the security and governance functions of most organisations were not built to match.

2. The governance gap

The clearest measure of the gap between adoption and governance comes from ISACA, whose 2026 AI Pulse Poll gathered responses from more than 3,400 digital trust professionals worldwide. It found that 90% believe employees in their organisation use AI, but that only 38% of organisations have a formal, comprehensive AI policy. That is an improvement on the 28% recorded a year earlier, and a further 30% have limited policies, but a quarter have no active policy at all. A third train all employees on AI, up from 22%. More telling for security are the operational findings: 56% of respondents do not know how quickly their organisation could halt an AI system in a security incident, 39% have no documented shutdown process, and only 43% are confident they could investigate and explain a serious AI incident to a regulator [16].

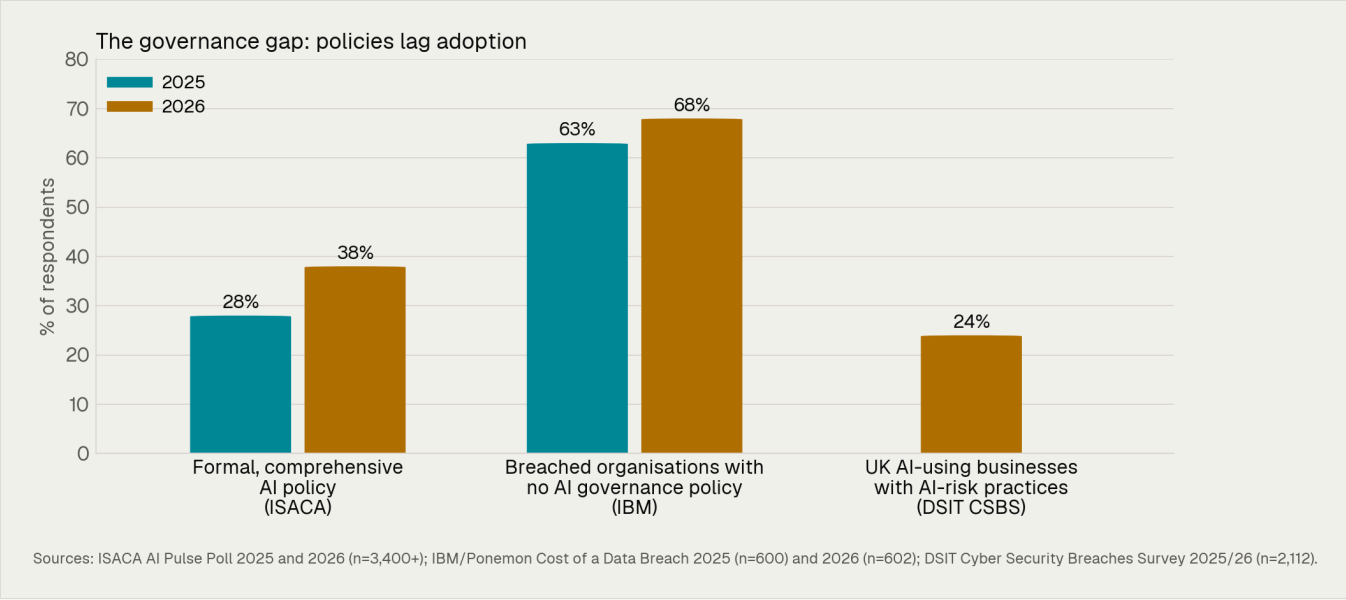


Figure. The governance gap: adoption against formal policy and controls. Sources: ISACA 2026 AI Pulse Poll [16]; IBM Cost of a Data Breach 2025 and 2026 [12][13]; DSIT Cyber Security Breaches Survey 2025/26 [11].

The UK picture is worse. The Cyber Security Breaches Survey 2025/26, the government's official statistics based on 2,112 businesses and 1,085 charities surveyed between August and December 2025, asked about AI for the first time. It found that 31% of businesses and 25% of charities were using AI, adopting it or actively considering it. Of that group, only 24% of businesses and 27% of charities reported having any cyber security practices or processes in place to manage the risks from

AI [11]. Read together, those two figures mean that roughly one UK business in thirteen has both adopted AI and thought about securing it. The same survey found the broader controls in retreat: the share of businesses carrying out cyber risk assessments fell from 48% to 41% in a year, and the share with formal cyber security policies fell from 59% to 52% [11].

IBM's Cost of a Data Breach research, which is built on interviews with organisations that have actually been breached rather than on self-assessment, provides the sharpest evidence of what the gap costs. Its 2025 edition found that 63% of breached organisations either had no AI governance policy or were still developing one [13]. The 2026 edition, drawing on 602 breaches between March 2025 and February 2026, puts the figure at 68% [12]. Among the organisations in the 2026 study that suffered a breach involving their own AI models or applications, 92% lacked adequate access controls on those systems; the corresponding figure a year earlier was 97% [12][13]. Governance is improving at the margin and remains absent in the majority.

> **UK CYBER DEFENCE VIEWPOINT: THE POLICY NOBODY WROTE**
>
> In our client work the pattern is consistent across sizes and sectors. Somebody in the business starts using a generative AI tool, usually for something harmless, and it works. Others follow. Six months later the finance team is pasting supplier invoices into a chatbot to reconcile them and the sales team has connected an AI note-taker to every customer call, and nobody in IT or security was asked because there was nothing to ask about: no purchase went through procurement and no system was installed. The organisation now has an AI programme, but it has no AI policy, no inventory of what is being used, and no idea what data has left.
>
> The remedy does not require a governance committee. It requires a page. The single most effective document we have seen a small organisation produce is a one-page AI use policy that says which tools are approved, which categories of data may never be put into any AI tool, who to ask when unsure, and what happens if it goes wrong, with a short list of the tools currently in use attached. ISACA's finding that only 33% of organisations train all staff on AI is, in our experience, the more fixable half of the problem, because the training is mostly a matter of telling people that the page exists.

3. Shadow AI and where the data goes

"Shadow AI" is the use of AI tools that the organisation has not sanctioned and cannot see. Three sources with quite different vantage points now measure it, and they agree that it is pervasive, that it is changing shape, and that it is where a great deal of sensitive data is leaving.

Verizon's 2026 Data Breach Investigations Report, which analysed more than 31,000 security incidents and 22,000 confirmed breaches, found that 45% of employees are now regular users of AI on their corporate devices, authorised or not, up from 15% the previous year, and that 67% of those users access AI services through non-corporate accounts. Shadow AI has become the third most common non-malicious insider action Verizon detects, a fourfold increase, and the most common type of data uploaded to external AI tools is source code [21]. A year earlier the same report had put routine access to generative AI on corporate devices at 15%, with 72% of it via personal email [22].

Netskope's cloud security telemetry, which observes what actually crosses the network, gives the most granular view. Its 2026 Cloud and Threat Report found that the share of AI users relying on personal AI applications fell from 78% to 47% over the year, while the share using organisation-managed accounts rose from 25% to 62%, as employers rolled out sanctioned tools; the share of people switching between personal and enterprise accounts more than doubled, from 4% to 9%. Data policy violations involving generative AI doubled over the same period. On average, 3% of an organisation's generative AI users commit violations, but each of them commits 223 a month. The data involved is source code in 42% of cases, regulated data such as personal or financial information in 32%, and intellectual property in 16%, with passwords and keys making up much of the remainder [14]. The average organisation now uses eight AI applications, up from six, and blocks ten; ChatGPT is in use at 77% of organisations, Google Gemini at 69% and Microsoft 365 Copilot at 52% [14]. Netskope's previous edition had found generative AI applications in use at 90% of organisations, with 72% of users on personal accounts and data traffic to generative AI up more than thirty-fold in a year [15].

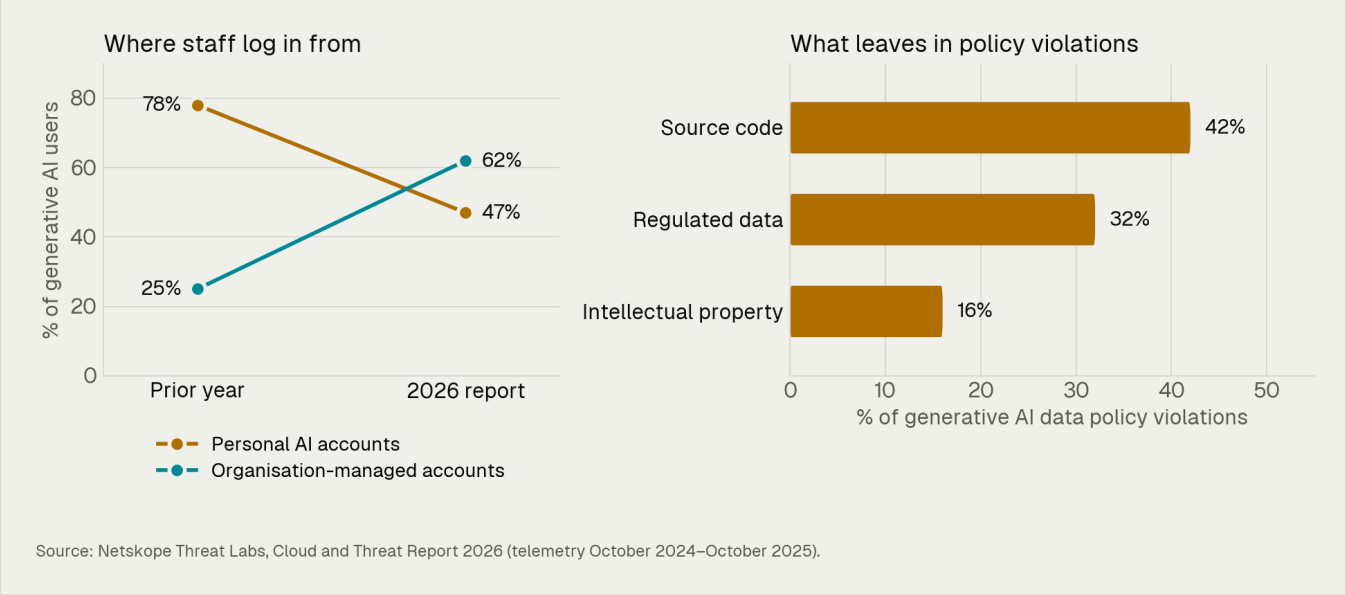


Figure. Where the data goes: types of data in generative AI policy violations, and the shift from personal to managed accounts. Source: Netskope Cloud and Threat Report 2026 [14].

Check Point's research, published in its 2026 AI Security Report and based on prompts observed across its customer base, measures the risk per interaction. Between January and May 2026, 3.45% of generative AI prompts worldwide, one in every 29, were classified as high risk because they contained sensitive data; the report describes this as a doubling over the year, from roughly one high-risk prompt in every fifty. The rate rose through the period, reaching 4.15%, or one in 24, in May. Europe had the highest regional rate at 3.95%, one in every 25 prompts, ahead of Latin America at 3.76%, North America at 3.33% and Asia-Pacific at 2.88%. Between 87% and 93% of organisations recorded at least one high-risk prompt in any given month. By industry, business services had the highest rate at 5.91%, or one prompt in seventeen, followed by wholesale and distribution at 5.47% and telecommunications at 4.06%; financial services, at 2.72%, and energy and utilities, at 1.91%, were well below the average, and information technology, at 0.65%, was lowest of all. The share of prompts going to external AI services rose from 2% to 4% over the period, and the average organisation was found to be running ten AI applications a month, many without formal approval [2].

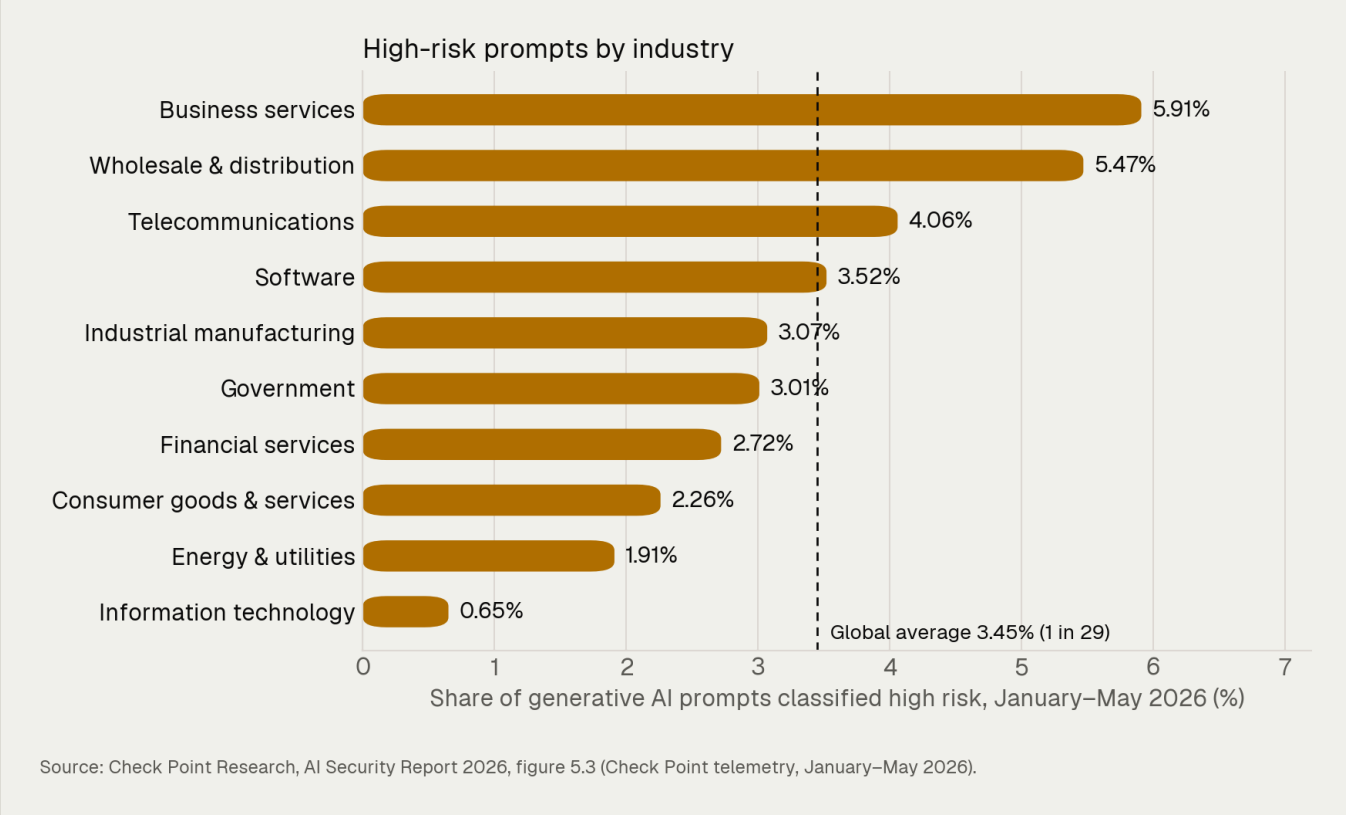


Figure. High-risk generative AI prompts by industry, January to May 2026. Source: Check Point Research, AI Security Report 2026 [2].

Gartner's survey of 302 security leaders adds the employer's perspective: 69% suspect or have evidence that employees are using prohibited public generative AI tools, and Gartner predicts that by 2030 more than 40% of enterprises will experience security or compliance incidents linked to unauthorised shadow AI [17]. A separate Gartner survey of employees found 57% using personal generative AI accounts for work and 33% entering sensitive data into them [19]. Microsoft's 2024 Work Trend Index, the origin of the widely quoted "bring your own AI" figure, found that 78% of AI users were bringing their own tools to work and that the figure was higher still, at 80%, in small and medium-sized companies; the two later editions of that report did not repeat the question, so the number should be cited as a 2024 finding [20].

What this costs is now measurable. IBM's 2025 study found that one in five breached organisations had suffered a breach linked to shadow AI, and that organisations with high levels of shadow AI paid on average US$670,000 more per breach than those with little or none; in such incidents customer personal data was compromised in 65% of cases, against 53% overall [13]. The 2026 edition reports shadow AI as a factor in 43% of breaches, more than double the prior year [12].

The incidents that illustrate the mechanism are well documented and predate most current tooling. In March 2023, within about three weeks of Samsung permitting its semiconductor engineers to use ChatGPT, three separate employees had pasted proprietary source code and an internal meeting transcript into it; the company restricted prompt sizes and, on 2 May 2023, banned generative AI tools on its devices and networks [99]. In January 2025, researchers at Wiz found a publicly accessible database belonging to the AI provider DeepSeek within minutes of examining its infrastructure, containing over a million lines of logs including chat histories, API keys and backend details, which the company secured on notification [97]. The first shows how easily data leaves through a sanctioned tool used carelessly; the second shows that the tool's own operator may be no better at protecting it.

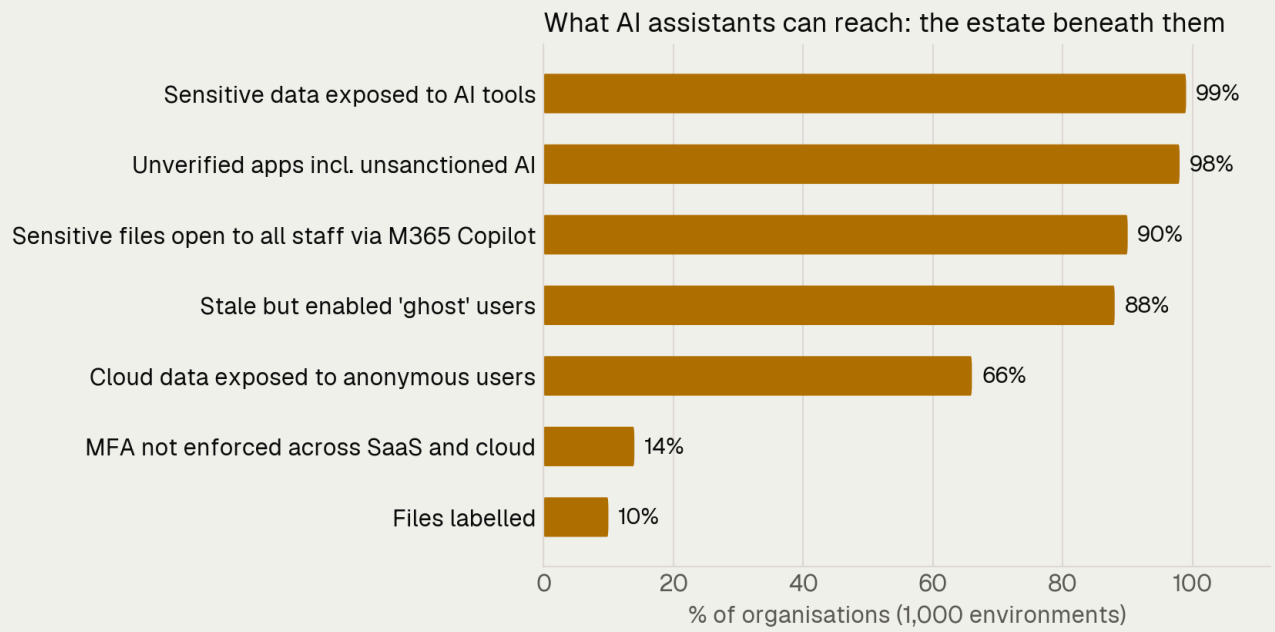
UK CYBER DEFENCE VIEWPOINT: BLOCK LESS, SEE MORE

The instinct when confronted with these figures is to block. Netskope's data shows that the average organisation now blocks ten AI applications, and the two most blocked are the ones the press has written most about. Blocking has its place, particularly for tools whose data handling terms are unacceptable, but as a strategy it fails for the same reason it failed against consumer cloud storage a decade ago: people route around it, and every route they take is one you cannot see. The organisations we see managing this well have done three things. They have given staff a sanctioned tool that is good enough that personal accounts stop being worth the bother, which is precisely the shift Netskope measured this year. They have put data loss prevention at the point of the prompt, so that a pasted customer list or credential is caught whether the destination is approved or not. And they have accepted that some use will always be unsanctioned and have made the consequences survivable, by keeping the credentials and data that matter most out of reach of the browser in the first place.

4. What your AI assistants can see

The largest AI risk in most organisations is not a new tool but an old problem made suddenly visible. For twenty years, the permissions on file shares, document libraries and cloud storage have been set by accident and never reviewed, and it has not mattered much, because finding a sensitive file required knowing where to look. An AI assistant connected to that estate removes the requirement. It will find, summarise and hand over anything the user's account can technically read, whether or not the user, or the organisation, knew the account could read it.

Varonis quantified this by examining the actual cloud environments of a thousand organisations, comprising nearly ten billion files, objects and records across the major SaaS and infrastructure platforms. It found sensitive data exposed to AI tools in 99% of organisations. In 90%, sensitive files were exposed to all employees through Microsoft 365 Copilot, and the average organisation had more than 25,000 sensitive folders open to every employee. Only one organisation in ten had labelled its files, without which data loss prevention and encryption controls cannot operate. Six per cent had sensitive files open to the internet, and 66% had cloud data exposed to anonymous users. The identity findings underneath were as poor: 88% of organisations had stale but still-enabled "ghost" user accounts, averaging 15,000 per organisation, with ten stale accounts holding administrative roles in the average estate; 98% had unverified applications, including unsanctioned AI, with 1,200 unofficial applications per company on average and one in four unverified OAuth applications rated high risk; and one organisation in seven did not enforce multi-factor authentication across its SaaS and cloud environments [1]. In Salesforce estates specifically, every organisation examined had at least one account able to export all its data, and one account in ten could do so freely, a permission that an Agentforce agent inherits when it acts on that account's behalf [1].



Source: Varonis, 2025 State of Data Security Report (1,000 organisations, ~10 billion cloud resources). 'MFA not enforced' shown as 14% (1 in 7).

Figure. Exposure of sensitive data to AI assistants and the identity hygiene beneath it. Source: Varonis, 2025 State of Data Security Report, 1,000 organisations [1].

IBM's 2026 findings point the same way from the breach side. Only 37% of breached organisations encrypted their sensitive data both at rest and in transit [12]. Orca Security's analysis of production cloud estates found that between 87% and 98% of organisations using AI services had never configured customer-managed encryption keys for them, rising to 98.4% for Amazon SageMaker notebooks, so that training data, models and pipelines rely entirely on the provider's default encryption [3].

The mechanism by which an assistant turns overexposure into loss is not hypothetical. The EchoLeak vulnerability in Microsoft 365 Copilot, disclosed and fixed in June 2025, allowed an attacker to extract data the assistant could reach by sending the target a single crafted email, with no click required; it is described in chapter 10 [64]. The important point for this chapter is that the vulnerability's reach was set entirely by the permissions already in place. An assistant that can see 25,000 sensitive folders is a much more valuable target than one that can see twelve.

> **UK CYBER DEFENCE VIEWPOINT: COPILOT READINESS IS PERMISSIONS WORK**
>
> We are asked, more than any other AI question, whether it is safe to turn on Microsoft 365 Copilot. The honest answer is that the question is the wrong way round. Copilot is safe in an estate where the permissions are right and dangerous in one where they are not, and it is the second kind that we usually find. Before enabling any assistant with access to organisational data, we advise three pieces of work, none of them glamorous: find and fix the sensitive locations that are open to "everyone" or to all staff, which Varonis's figures suggest will be measured in the tens of thousands of folders in a mid-sized firm; disable the stale accounts, especially the administrative ones, that would give an attacker a quiet foothold from which to ask the assistant questions; and label the data categories that must never be summarised, so that the platform's own controls can enforce it. Done properly, that work takes weeks rather than days, and it is worth doing whether or not Copilot follows, because it is the same work that limits the blast radius of any compromised account.

PART

Part two: the threat, as the evidence shows it

5. What the official assessments say

The most authoritative statements about how AI is changing the cyber threat to the United Kingdom come from the National Cyber Security Centre, and they are worth quoting precisely, because they are more measured than most of what is written about them.

The NCSC's first assessment, published in January 2024, judged that AI "will almost certainly increase the volume and heighten the impact of cyber attacks over the next two years", but that the threat to 2025 would come "from evolution and enhancement of existing tactics, techniques and procedures". It found that "all types of cyber threat actor, state and non-state, skilled and less skilled, are already using AI, to varying degrees", that AI "provides capability uplift in reconnaissance and social engineering, almost certainly making both more effective, efficient, and harder to detect", and that it "lowers the barrier for novice cyber criminals, hackers-for-hire and hacktivists to carry out effective access and information gathering operations". Advanced use of AI in malware and exploit development was judged "highly likely to be restricted" to well-resourced actors and "unlikely to be realised before 2025" [23]. The NCSC's chief executive at the time summarised the finding as "evolutionary not revolutionary".

The follow-up assessment to 2027, published in May 2025, is firmer. It judges that AI "will almost certainly continue to make elements of cyber intrusion operations more effective and efficient, leading to an increase in frequency and intensity of cyber threats"; that "there will almost certainly be a digital divide between systems keeping pace with AI-enabled threats and a large proportion that are more vulnerable, making cyber security at scale increasingly important to 2027 and beyond"; that "proliferation of AI-enabled cyber tools will highly likely expand access to AI-enabled intrusion capability to an expanded range of state and non-state actors"; and that the growing incorporation of AI systems across the UK's technology base, particularly in critical national infrastructure, "almost certainly presents an increased attack surface". It states that the time between vulnerability disclosure and exploitation "has shrunk to days and AI will almost certainly reduce this further". Its most-quoted judgement is conditional and should be quoted with its condition: "Assuming a lag, or no change to cyber security mitigations, there is a realistic possibility of critical systems becoming more vulnerable to advanced threat actors by 2027" [24].

The NCSC's Annual Review 2025, covering the year to August 2025, records that the Centre handled 429 incidents, of which 204, or 48%, were nationally significant, more than double the 89 of the previous year, with eighteen highly significant incidents against twelve. On AI it states that "threat actors of all types continue to use AI to enhance their existing tactics, techniques and procedures, rather than to create novel attacks", that actors linked to China, Russia, Iran and North Korea "are using large language models to evade detection, support reconnaissance, process exfiltrated data, access systems through social engineering, and support vulnerability research and exploit development", and that "the most significant AI-cyber development in the near-term will highly likely come from AI-assisted" vulnerability research and exploit development [25].

The most striking recent evidence of capability comes from the UK's AI Security Institute, whose evaluations the NCSC cited in a joint blog in March 2026. On a simulated 32-step attack against a realistic enterprise network, the best frontier model completed an average of 15.6 steps, and at its best 22 of 32, where eighteen months earlier the best models had barely managed two. A full attempt cost around £65 and took the model roughly six of the fourteen hours a skilled human would

need. The NCSC's conclusion was that "AI won't compensate for weak security foundations, but it will amplify both strengths and weaknesses", and that defenders "should assume that at least some attackers already have access to capable AI tools" [26]. In May 2026 the Institute reported that the length of cyber task the best models can complete autonomously has been doubling roughly every 4.7 months since late 2024, faster than its previous estimate of eight months, and noted that "cyber defenders have reported significant advances in vulnerability discovery using recent models" [27].

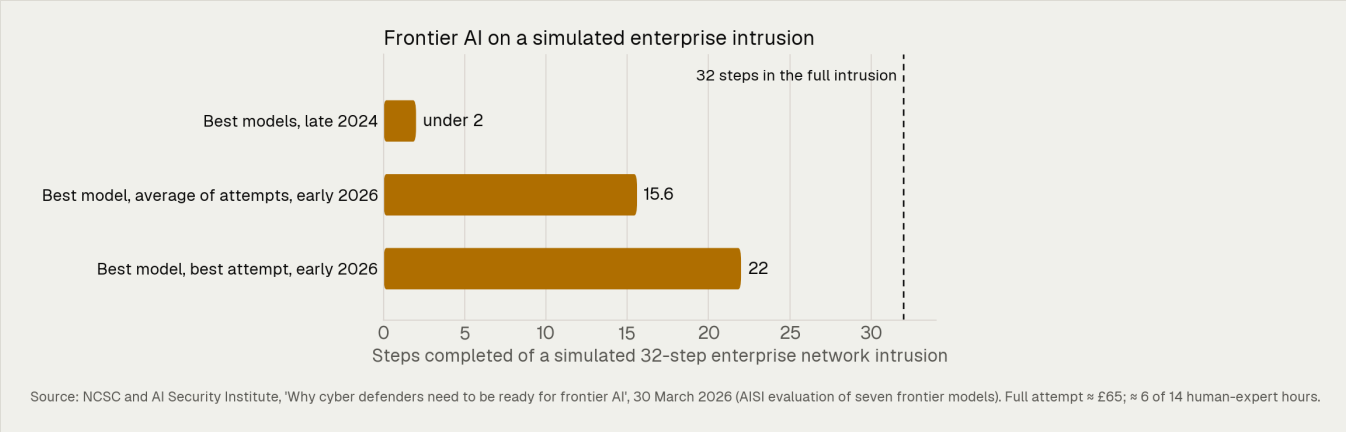


Figure. Frontier AI on a simulated 32-step enterprise intrusion: best model's completed steps, late 2024 to early 2026. Source: NCSC/AISI, March 2026 [26].

6. AI-enabled attacks in the breach data

Assessments describe what is possible. Breach studies describe what happened. Four large ones published in 2026 give a consistent and somewhat sobering picture, in which AI is now a routine factor in attacks without having transformed their nature.

IBM's Cost of a Data Breach 2026, based on 602 organisations breached between March 2025 and February 2026, found that one in four malicious breaches was AI-enabled, a 56% increase over the previous year, and that these breaches cost on average US$6 million, roughly a million dollars more than the global average of US$4.99 million, itself a record high and 12% above the year before. The most common forms were deepfake impersonation and AI-enabled malware. Among UK organisations, 22% reported experiencing AI-generated attacks, of which deepfake impersonation accounted for 45%, AI-enabled malware 19% and phishing 17%; the average UK breach cost £3.13 million, slightly down from £3.29 million [12]. A year earlier IBM had found attackers using AI tools in 16% of breaches [13].

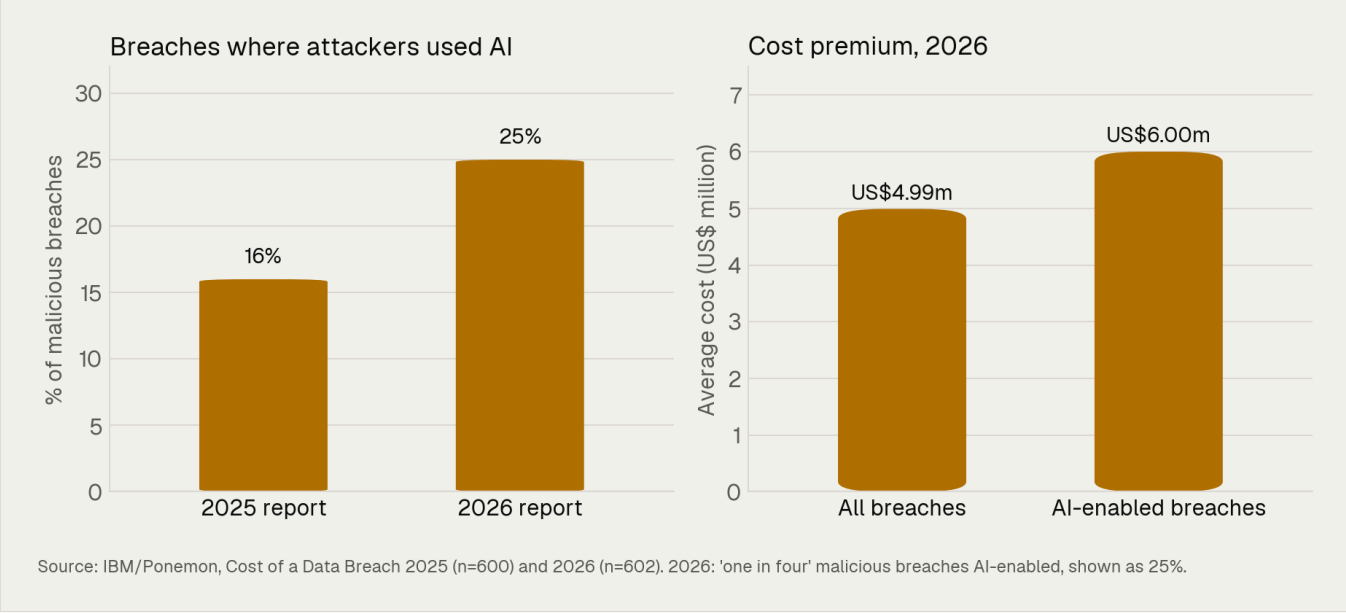


Figure. AI in the breach data: share of malicious breaches that were AI-enabled and their cost premium. Source: IBM/Ponemon, Cost of a Data Breach 2025 and 2026 [12][13].

Verizon's 2026 Data Breach Investigations Report, the largest study of its kind with more than 31,000 incidents and 22,000 confirmed breaches from 145 countries, found the human element present in 62% of breaches. Its analysis of what threat actors actually do with AI, drawn from their queries to AI platforms, found that a typical actor had researched or used AI assistance across fifteen documented techniques, some as many as forty or fifty, and that phishing accounted for 44% of the AI-assisted initial access activity it observed. Its conclusion is precise: "most AI-assisted development of malware and tooling was associated with well-known and defined attack techniques", fewer than 2.5% of AI-assisted malware observations involved uncommon techniques, and the incident data showed no meaningful increase in phishing success despite the assistance. The larger shifts in the 2026 data were elsewhere: exploitation of vulnerabilities became the leading initial access vector at 31% of breaches, up from 20%, credential abuse fell from 22% to 13%, phishing held at 16%, third parties were involved in 48% of breaches, up from 30%, and ransomware was present in 48% [21][22].

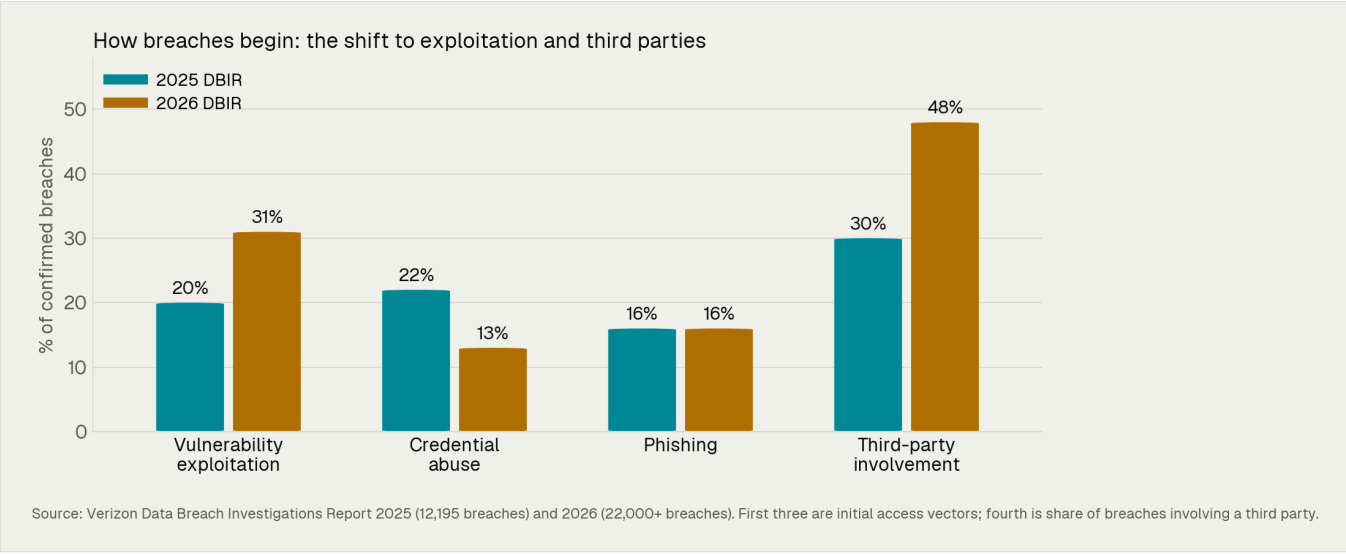


Figure. Initial access in confirmed breaches, 2025 and 2026 reports. Source: Verizon DBIR 2025 and 2026 [21][22].

Google's Mandiant, in M-Trends 2026, drawing on more than 500,000 hours of incident investigation in 2025, is equally direct: "we do not consider 2025 to be the year where breaches were the direct result of AI". It did, however, observe malware families that query large language models during execution to evade detection, and a species of malware that checks victims' machines for locally installed AI command-line tools, a sign of where attackers expect value to lie. Exploits were the leading initial infection vector for the sixth year at 32% of intrusions, voice phishing second at 11%, and email

phishing accounted for only 6%. The speed statistic that matters most is not about AI directly: the median time between an initial compromise and its hand-off to a second criminal group fell from more than eight hours in 2022 to 22 seconds in 2025. Median dwell time, the period an attacker is present before detection, rose to fourteen days from eleven [30].

Microsoft's Digital Defense Report 2025 found that more than half of attacks with a known motive were financial, that identity-based attacks rose 32% in the first half of 2025 and that more than 97% of them were password spray or brute-force attempts, which phishing-resistant multi-factor authentication blocks. It reports adversaries using generative AI "for a variety of activities, including scaling social engineering, automating lateral movement, engaging in vulnerability discovery", and prompt injection and data poisoning "to attack both models and systems" [28]. A Microsoft analysis in April 2026 added that a single adversary-in-the-middle phishing platform, Tycoon2FA, accounted for roughly 62% of all phishing attempts Microsoft was blocking each month and had been linked to nearly 100,000 compromised organisations since 2023, the UK being its second most targeted country [29]. ENISA's Threat Landscape 2025, based on 4,875 incidents, concludes that AI functions "as an optimisation tool for malicious activities but also as a new point of exposure", that language models "are being used to enhance phishing and automate social engineering", and that "attacks on the AI supply chain are on the rise" [31].

> **UK CYBER DEFENCE VIEWPOINT: THE ATTACKS WE SEE ARE THE OLD ONES, DONE BETTER**
>
> The breach studies match what our own operations centre sees. The intrusions that reach our clients are overwhelmingly the familiar kinds: a phished session, an exposed service, a supplier's credentials, a vulnerability exploited days after disclosure. What has changed is the quality and the tempo. Phishing emails no longer read as if translated; they reference the right project and the right colleague. Exploitation follows disclosure in days rather than weeks, which matches the NCSC's judgement exactly. And the adversary-in-the-middle phishing kits that now dominate, which proxy the genuine Microsoft sign-in page and steal the session after the victim has completed multi-factor authentication, are sold as a service and need no skill to run; we published a walk-through of one such attack against a client in September 2026 [100]. None of this is a reason to buy an AI product. It is a reason to fix the fundamentals faster, because the window between a weakness appearing and its being used has closed to the point where a monthly patch cycle is no longer a defence.

7. Phishing, deepfakes and fraud

The single question business leaders most often ask about AI security is whether AI-written phishing works better. The evidence is more interesting than a yes or no, and it has been widely misreported.

The figure most often quoted, that AI phishing achieves a 54% click rate against 12% for ordinary phishing, a 4.5-fold increase, comes from a Harvard study by Heiding, Lermen, Kao, Schneier and Vishwanath, first published in late 2024 and peer-reviewed in Expert Systems with Applications in June 2026. It is a rigorous experiment, but it involved 101 participants. The fully automated AI campaign and the human expert campaign each achieved 54%, AI with a human in the loop achieved 56%, and the control 12%; the authors' point was that AI matched expert humans at a fraction of the cost, increasing potential profitability "by up to 50 times" [40]. Microsoft's Digital Defense Report cites these figures, and they have since been repeated as if they were Microsoft telemetry, which they are not [28].

Two larger studies complicate the picture. Hoxhunt, running roughly 70,000 simulations across 2.5 million users in each of two waves, found its AI spear-phishing agent had gone from 31% less effective than elite human red teams in March 2023 to 24% more effective in March 2025, though the absolute failure rates in question were low, at 2.78% for AI against 2.25% for humans, and the methodology changed between waves [41]. A 2026 field experiment at a German university with 7,700 participants found that personalised phishing written by a language model achieved a 10.0% click rate against 3.7% for generic AI phishing, almost tripling it at a cost of about three cents an email, but that personalised phishing written by humans still did better, at 24.2% [42]. The fair summary is that AI has made good phishing cheap and plentiful rather than uniquely persuasive, and that the best human attackers remain more effective than the best automated ones for now.

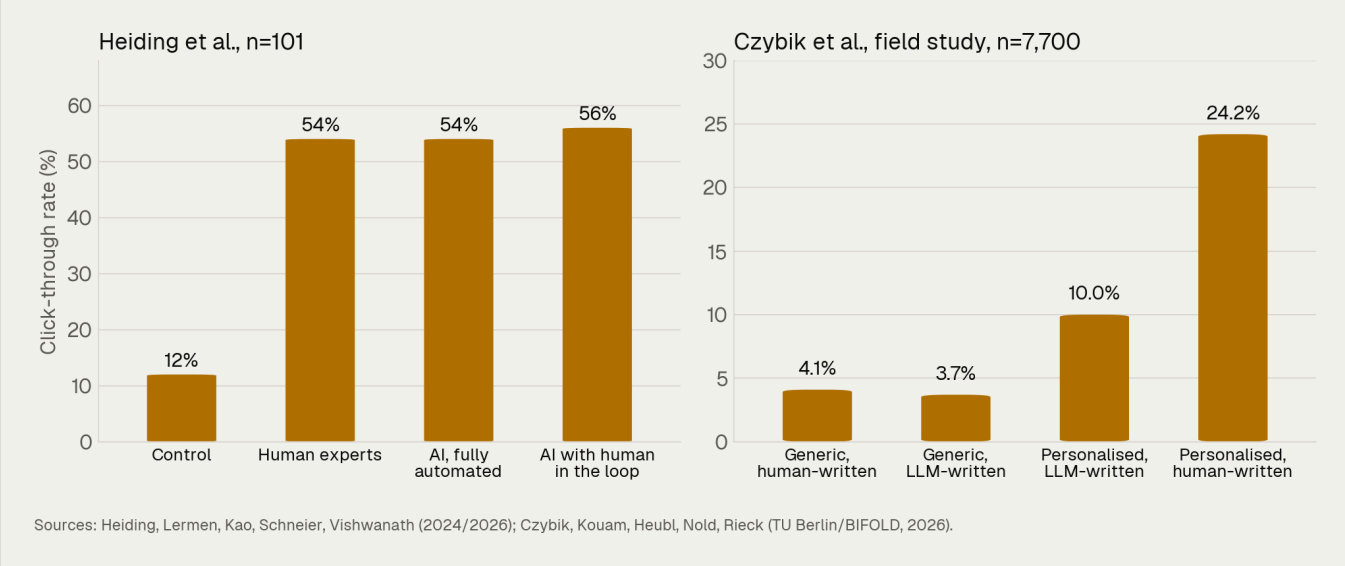


Figure. What the studies actually found: click-through rates in three experiments on AI-written phishing. Sources: Heiding et al. (n=101) [40]; Czybik et al. (n=7,700) [42].

Whether AI phishing works better matters less than whether it is being caught, and here the data is clear that it is not being caught in volume. Verizon found that synthetically generated text in malicious emails had doubled over two years to roughly a tenth of the total [22]. Hoxhunt itself noted that only 0.7% to 4.7% of the phishing that bypassed email filters in 2024 was AI-written [41]. A figure sometimes attributed to ENISA, that more than 80% of social engineering is now AI-supported, is a vendor estimate that ENISA relays with the word “reportedly”, and we do not rely on it [31].

Deepfake fraud has moved from anecdote to statistics. The case that established it involved an employee of the engineering firm Arup in Hong Kong, who in early 2024 joined a video conference in which the chief financial officer and several colleagues were deepfaked and made fifteen transfers totalling HK$200 million, about US$25 million; Arup confirmed that “fake voices and images were used” and that its own systems were not compromised [46]. The FBI’s 2025 Internet Crime Report, published in April 2026, tracked AI-enabled fraud for the first time and recorded 22,364 complaints and nearly US$893 million in losses, from a total of just over a million complaints and nearly US$21 billion in losses; because there is no prior year, no growth rate can honestly be derived from it [43]. Gartner’s survey of security leaders found 62% had experienced a deepfake attack in the previous year [18]. Check Point’s research found that ordinary viewers correctly identified only about 41% of AI-generated faces, and in one test only about 30% [2].

In the United Kingdom, the City of London Police’s first annual fraud assessment, published in September 2026, recorded that reports identifying AI as a factor rose from 193 in 2024/25 to 956 in 2025/26, an increase of 395%, and that losses linked to AI-enabled crime rose from £1.2 million to £9.6 million, with investment fraud the largest category and criminals using cloned voices to defeat telephone banking security. The police caution that victims are often unaware that AI was used, so these figures undercount [44]. UK Finance’s 2026 report put total fraud losses in 2025 at £1.28 billion, with authorised push payment losses up 19% to £576 million, and describes criminals “exploiting advances like AI to industrialise operations”, without an AI-specific figure [45]. Deloitte’s much-cited forecast that generative AI could push US fraud losses to US$40 billion by 2027 is a scenario projection from a 2023 baseline of US$12.3 billion, not a measurement, and applies to the United States only [47].

> **UK CYBER DEFENCE VIEWPOINT: THE DEFENCE IS THE PROCESS, NOT THE DETECTION**
>
> Every deepfake incident we have investigated or read about was defeatable by a process that had nothing to do with recognising a fake. A payment of any size to a new or changed account is confirmed by a call to a number already on file, never by replying to the request. An urgent instruction from a senior person by video or voice is confirmed by a second channel before it is acted on, and the person is told in advance that this will happen so that it does not feel like distrust. Two people are needed to move money above a threshold. These controls cost nothing, they are already required by most audit standards, and they hold regardless of how convincing the voice on the line becomes. Training staff to spot deepfakes is worth doing, but Check Point's finding that ordinary people catch fewer than half of them tells you not to rely on it. The good morning here is the one where the finance clerk says, politely, that they will call back.

8. What the AI companies are seeing

A distinctive feature of 2025 and 2026 has been the AI developers themselves publishing what they see on their platforms. Their reports are unusually valuable because they observe attackers at the moment of using the tool, and unusually easy to over-read.

Anthropic's August 2025 report stated that "agentic AI has been weaponized", that AI "has lowered the barriers to sophisticated cybercrime", and that criminals "have embedded AI throughout all stages of their operations". It documented an extortion operation in which a single actor used an AI agent to attack at least seventeen organisations including healthcare providers and government bodies, with ransom demands sometimes exceeding US$500,000, and North Korean operatives using AI to obtain and hold remote technology jobs [32]. In November 2025 it reported disrupting what it described as the first AI-orchestrated cyber espionage campaign, attributed with high confidence to a Chinese state-sponsored group, in which the AI performed "80 to 90% of the campaign, with human intervention required only sporadically" against roughly thirty targets, succeeding against a small number. The same report is candid that the AI "occasionally hallucinated credentials or claimed to have extracted secret information that was in fact publicly available", which "remains an obstacle to fully autonomous cyberattacks" [33]. Its September 2026 report, covering December 2025 to August 2026, concludes that "sophisticated attacks no longer require sophisticated attackers", that AI "has collapsed the labour and tooling gap that used to separate well-resourced, state-sponsored operations from individual operators", and that a majority of the operations it describes were enabled by AI through direct execution or orchestration; it documents intrusions that exfiltrated more than 300,000 national identity records and more than a terabyte of data, and notes that stolen API keys are now traded as loot, compute and cover [34].

OpenAI has published a series of disruption reports since February 2024 and had reported more than forty disrupted networks by October 2025. Its consistent finding is that threat actors "bolt AI onto old playbooks to move faster, not gain novel offensive capability", building AI into existing workflows rather than new workflows around AI; it also estimated that its chatbot was being used to identify scams up to three times more often than to perpetrate them [35]. Its 2026 reports concern chiefly influence operations and scams, and several statistics attributed to them in secondary reporting do not appear in the reports themselves.

Google's Threat Intelligence Group has tracked the trajectory most systematically. In January 2025 it saw no indication of threat actors "developing novel capabilities" and judged that current models "on their own are unlikely to enable breakthrough capabilities" but did let actors "move faster and at higher volume" [36]. By November 2025 it reported, "for the first time", malware families that use language models during execution [37]. In May 2026 it reported, again for the first time, "a threat actor using a zero-day exploit that we believe was developed with AI", intended for mass exploitation and pre-empted by Google's own discovery [38]. In September 2026 it described adversaries moving "from basic prompting to agentic AI workflows", including a case in which a criminal compromised a cloud resource and then planned, built and executed an agent-enabled mass credential-harvesting campaign in under six hours; it also reported model-extraction campaigns of more than a hundred million prompts. Its bottom line remains that it "has not yet observed threat actors deploying fully autonomous pipelines against targets in the wild" [39].

Check Point's 2026 report, drawing on its own incident work, describes intrusions in which AI "ran exploitation workflows autonomously, generating thousands of commands across dozens of sessions with minimal human direction", malware "that a single developer produced in under a week at a quality level our researchers initially attributed to a multi-person team working for months", and an observation that ought to concern every defender: in most of these cases "what gave away the AI's role in the attack was the attacker's own operational mistakes or monitoring by the AI provider, not anything the victim organisation had detected". It also notes that attackers obtain AI capability mostly by abusing commercial models with legitimate or stolen accounts, that self-hosted models remain "more aspirational than practical" for them, and that demand for purpose-built criminal AI services "peaked and then declined as the underground grew sceptical of their quality" [2].

Read together, these reports describe a capability that is rising steeply, that is already doing most of the work in some real intrusions, and that has not yet produced an attack that runs from start to finish without a person. That is a description of a trajectory, not a plateau, and the UK's own AI Security Institute has measured the rate of climb. The sensible planning assumption is the NCSC's: that at least some of the people attacking you already have these tools.

PART

Part three: securing the AI you deploy

9. The attack surface of an AI system

An organisation that deploys an AI system, whether a chatbot on its website, an assistant over its documents, a model it has fine-tuned, or an agent that acts on its behalf, has added a new class of application with a new class of weakness. The good news is that the weaknesses are now well catalogued, and the catalogues agree.

The OWASP Top 10 for Large Language Model Applications, in its 2025 edition, is the reference list for the risks specific to systems built on language models. Its ten entries are prompt injection, sensitive information disclosure, supply chain, data and model poisoning, improper output handling, excessive agency, system prompt leakage, vector and embedding weaknesses, misinformation, and unbounded consumption [48]. In December 2025 OWASP added a companion list for agentic systems, whose ten entries are agent goal hijack, tool misuse and exploitation, identity and privilege abuse, agentic supply chain vulnerabilities, unexpected code execution, memory and context poisoning, insecure inter-agent communication, cascading failures, human-agent trust exploitation, and rogue agents [49]. Both lists are set out in full, with the controls that address each entry, in annex A.

OWASP TOP 10 FOR LLM APPLICATIONS (2025)	OWASP TOP 10 FOR AGENTIC APPLICATIONS (2026)
LLM01 Prompt Injection	ASI01 Agent Goal Hijack
LLM02 Sensitive Information Disclosure	ASI02 Tool Misuse and Exploitation
LLM03 Supply Chain	ASI03 Identity and Privilege Abuse
LLM04 Data and Model Poisoning	ASI04 Agentic Supply Chain Vulnerabilities
LLM05 Improper Output Handling	ASI05 Unexpected Code Execution
LLM06 Excessive Agency	ASI06 Memory and Context Poisoning
LLM07 System Prompt Leakage	ASI07 Insecure Inter-Agent Communication
LLM08 Vector and Embedding Weaknesses	ASI08 Cascading Failures
LLM09 Misinformation	ASI09 Human-Agent Trust Exploitation
LLM10 Unbounded Consumption	ASI10 Rogue Agents

Table 1. The two OWASP lists. Sources: OWASP GenAI Security Project [48][49].

MITRE ATLAS, the knowledge base of adversary tactics and techniques against AI systems modelled on the ATT&CK framework, is the reference for what attackers have actually done. Its September 2026 release describes sixteen tactics, 114 techniques and 83 sub-techniques, supported by 72 documented case studies and 39 mitigations; it is updated monthly and the August 2026 release alone added eighteen techniques concerning autonomous agent behaviour [50]. NIST's AI Risk Management Framework provides the governance structure, with its four functions of govern, map, measure and manage, and its Generative AI Profile of July 2024 identifies twelve risks specific to or exacerbated by generative AI, including

confabulation, data privacy, information security and value-chain integrity [51]. NIST's draft Cyber AI Profile, published for comment in December 2025, extends the Cybersecurity Framework to the three concerns of securing AI systems, defending with AI, and thwarting AI-enabled attacks [52].

For UK organisations the most directly applicable document is the government's Code of Practice for the Cyber Security of AI, published in January 2025, which sets out thirteen principles across the secure design, development, deployment, maintenance and end-of-life of AI systems, from raising awareness of AI security threats and evaluating them, through securing infrastructure and supply chain and documenting data, models and prompts, to monitoring the system's behaviour and disposing of data and models properly [54]. It became the European standard ETSI TS 104 223 in April 2025, with 72 specific provisions, and was elevated to the full European Norm EN 304 223 in December 2025 [55]. It builds on the guidelines for secure AI system development that the NCSC and CISA published in November 2023 with 21 other agencies from eighteen countries, which organise the task into secure design, secure development, secure deployment, and secure operation and maintenance [53]. ISO/IEC 42001, published in December 2023, provides the certifiable management-system standard for organisations that develop, provide or use AI; ISO has not yet reported certification numbers and the only available estimates put global certifications in the low hundreds by spring 2026 [56].

The frameworks are mapped to one another in annex B. The practical point is that an organisation does not need to choose between them. The UK code and its ETSI standard tell you what to do; the OWASP lists and MITRE ATLAS tell you what you are defending against; NIST tells you how to govern the whole; and ISO 42001 lets you prove it.

10. Prompt injection and data exfiltration in real products

Prompt injection is the first entry on the OWASP list because it is the weakness that distinguishes AI systems from everything that came before. A language model cannot reliably tell the difference between the instructions it was given and the content it is asked to process. If the content it processes, an email, a web page, a document, a calendar invitation, a support ticket, contains instructions, the model may follow them. When the model has access to data and to tools, following them can mean handing the data to an attacker. The attacker need never touch the system directly; they need only get their text in front of it.

The year's disclosures established that this is not theoretical, and that it affects the products organisations most commonly use. The following were all reported responsibly, fixed by the vendors, and are not known to have been exploited; they are set out here at the level of a security advisory, and in tabular form in annex C.

EchoLeak, tracked as CVE-2025-32711, was a zero-click vulnerability in Microsoft 365 Copilot disclosed by Aim Security and fixed by Microsoft in June 2025. A single crafted email, which the victim did not need to open, could cause the assistant to disclose data it could reach; Microsoft rated it critical at 9.3, NIST at 7.5, and no customer action was required because the fix was server-side [64]. ForcedLeak, disclosed by Noma Security in September 2025 and fixed by Salesforce, allowed data to be extracted from Agentforce through instructions placed in a web-to-lead form, aided by an expired domain that remained on the platform's trusted list [65]. In 2026, Varonis Threat Labs disclosed SearchLeak, CVE-2026-42824, a one-click flaw in Microsoft 365 Copilot's enterprise search mitigated by Microsoft in June; a further Copilot vulnerability, CVE-2026-24299, reported in October 2025 and fixed in March 2026, combined data exfiltration with persistence through the assistant's memory feature; and Zenity Labs disclosed AgentForger, a flaw in ChatGPT's workspace agents that OpenAI fixed within days in June 2026 [66]. Orca's researchers demonstrated, in a piece of work they call RoguePilot, a passive prompt-injection chain achieving full repository takeover through GitHub Copilot and Codespaces [3].

The pattern across these cases is instructive. In each, the attacker's text arrived through a legitimate channel the assistant was designed to read. In each, the damage was bounded by what the assistant was permitted to reach. And in each, detection by the victim organisation would have been difficult, because the assistant did exactly what an assistant does.

The survey evidence suggests such attacks are now common. Gartner's 2025 survey of 302 security leaders found that 29% had experienced an attack on their generative AI application infrastructure in the previous twelve months and 32% an attack on AI applications through prompt manipulation [18]. Check Point reports that the detection rate of malicious prompts across its customers rose through early 2026 to approach 1% of everything observed in May, and that in one class of attack, in which

a single legitimate-sounding request decomposes into individually benign steps, the attack succeeded in 71% of cases [2]. Gartner forecasts that by 2029 “over half of successful cyberattacks on AI agents are expected to exploit access control weaknesses and prompt injections” [77].

Data poisoning, the fourth OWASP entry, was given a firm empirical footing in October 2025 by a joint study from Anthropic, the UK AI Security Institute and the Alan Turing Institute. It found that “as few as 250 malicious documents can produce a backdoor vulnerability in a large language model, regardless of model size or training data volume”, across models from 600 million to 13 billion parameters; for the largest, 250 documents amounted to 0.00016% of the training data. The authors are careful to note that the backdoor they studied was narrow, causing the model to produce gibberish on a trigger, and that “it remains unclear how far this trend will hold” at larger scale [76]. For an organisation fine-tuning a model on its own data, or building retrieval over documents it does not fully control, the implication is that the integrity of training and reference data is a security control, not merely a quality one.

> **UK CYBER DEFENCE VIEWPOINT: AN AI ASSISTANT IS AN INTERNET-FACING APPLICATION**
>
> The mental model that serves best is to treat every AI assistant and agent with access to organisational data as though it were a web application exposed to the internet, because functionally it is. Anyone who can get text in front of it, by email, by web form, by a document in a shared drive, by a calendar invitation, is an unauthenticated user of it. That reframing produces the right controls almost automatically. You would not give an internet-facing application a service account with read access to every file in the company, so you do not give it to the assistant. You would log and monitor what an internet-facing application does, so you log what the assistant reads and what it sends. You would patch it promptly, which for a SaaS assistant means keeping the connectors and plug-ins current and removing the ones you do not use. And you would test it, which for an assistant means asking a competent tester to try to make it misbehave before an attacker does. We have started including AI assistants in scope for penetration tests by default, and the findings are rarely about the model; they are about what it was allowed to reach.

11. The AI supply chain

The third OWASP entry is supply chain, and the year’s evidence suggests it may be the one that hurts most organisations first, because it requires no sophistication on the attacker’s part and no AI system on the victim’s. It requires only that developers and staff use AI tools, models and packages that someone else made.

The models themselves can be malicious. Hugging Face, the principal public repository, reported in April 2025 that its scanning partner had examined 4.47 million model versions in 1.41 million repositories and found 352,000 unsafe or suspicious issues across 51,700 models [71]. JFrog had found around a hundred models on the platform in February 2024 that executed hidden code when loaded, one opening a reverse connection to a remote server, using the serialisation format that most machine-learning tooling still relies on; ReversingLabs in February 2025 found models using deliberately corrupted files to evade the platform’s scanners [72]. The practical rule is that a model file is executable code and should be handled as such: obtained from a trusted publisher, scanned, and loaded in safe formats only.

The packages that AI writes can be fictitious. A USENIX Security 2025 study by Spracklen and colleagues generated 576,000 code samples from sixteen language models and found that they recommended software packages that did not exist in at least 5.2% of cases for commercial models and 21.7% for open-source ones, producing 205,474 unique hallucinated package names, a large share of which recurred consistently [73]. An attacker who registers a package under one of those names has only to wait. The defence is the ordinary one of verifying that a dependency exists, is maintained, and is the one intended, applied with more rigour now that a proportion of suggestions are invented.

The AI tools developers use can be turned against them. In August 2025 a compromised build pipeline published malicious versions of the widely used Nx build system to the npm registry for a little over four hours. The malicious code harvested credentials from developers’ machines and, notably, invoked any locally installed AI coding assistants, instructing them to search the filesystem for secrets; it then published the results to repositories under the victims’ own accounts. Analyses found more than 2,300 distinct secrets exposed, more than a thousand valid GitHub tokens, and a second wave in which over 5,500 private repositories from more than 400 organisations were made public [74]. In July 2025 a version of Amazon’s Q

Developer extension for Visual Studio Code shipped with code, introduced by a threat actor through an inadequately scoped access token, that instructed the assistant to delete files and cloud resources; the code failed to execute because of a syntax error, and Amazon replaced the version within two days [75]. Check Point's researchers, scanning some 46,500 published code packages, found that 428 had accidentally included a developer's local AI-assistant settings file and that about one in thirteen of those contained live credentials, including registry, GitHub and Hugging Face tokens; separately, examining 221 pre-built skills in a public marketplace for one AI agent, they found 70% over-requested credentials and 43% contained command-injection patterns, and one campaign placed 44 malicious skills in the store that were downloaded more than 12,500 times [2].

The connectors through which agents reach tools are a supply chain of their own. The Model Context Protocol, now the common standard for connecting AI agents to data and tools, has its own published security best practices, and the year produced both a critical vulnerability and the first malicious server. CVE-2025-6514, in the widely used mcp-remote client, allowed a malicious server to execute commands on a connecting machine; it was rated 9.6 and fixed in July 2025 [67]. In September 2025 a package on npm posing as a connector for the Postmark email service was found to be silently copying every email sent through it to an external address; it had about 1,600 downloads, and the genuine company had never published such a package [68]. Check Point reports finding security weaknesses in 40% of 10,000 MCP servers it reviewed [2]. The NSA published design considerations for MCP security in May 2026 [70].

Orca's telemetry shows how far the hygiene around all of this has to travel. Among more than 1,200 production cloud estates, 81% of organisations with AI packages carried at least one known high-severity vulnerability and 74% at least one rated critical; 99.9% of fixable AI vulnerabilities remained unpatched, and the share of AI vulnerability alerts with a public exploit available rose from 0.2% to 50.1% year on year. Three in ten AI adopters had at least one AI secret stored insecurely, the share of OpenAI users storing keys in insecure locations having risen from 20% to 28%, and exposure of Anthropic keys having tripled. AI browser extensions were 60% more likely to carry known vulnerabilities than other extensions. Four in five organisations using Amazon SageMaker ran it with all five of its core insecure defaults enabled [3].

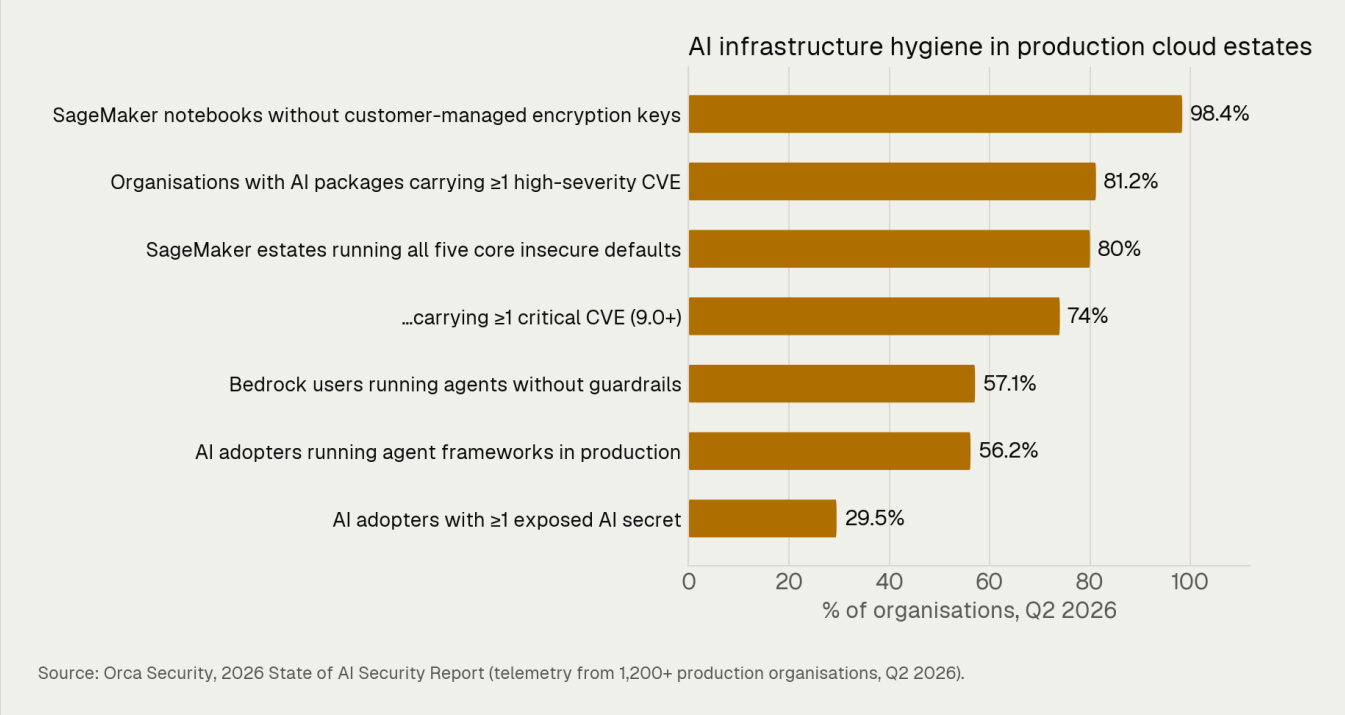


Figure. The state of AI infrastructure hygiene in production cloud estates, Q2 2026. Source: Orca Security, 2026 State of AI Security Report, 1,200+ organisations [3].

UK CYBER DEFENCE VIEWPOINT: THE DEVELOPER'S LAPTOP IS THE NEW PERIMETER

The Nx incident is the one we discuss most with clients, because it combined two things we already knew into something new. Developers keep credentials on their machines; that has always been a risk. Developers now also keep AI assistants on their machines that will, if asked, search for those credentials with great efficiency. The malware did not need to know where the secrets were. It asked. The lesson is not to stop using AI coding tools, which are now part of how software is made, but to stop keeping long-lived secrets where any process on the machine can read them, to run AI assistants with the least privilege the task needs and never with permission checks disabled, and to treat every AI tool, model, package and connector as a supplier: known, listed, updated, and removed when no longer used. That is the same supplier-register discipline we wrote about for cloud services in September, applied to a supply chain that most organisations have not yet noticed they have.

PART

Part four: agents, identities and containment

12. The agentic shift

An AI agent is a system that does not merely answer but acts: it is given a goal and the means to pursue it, and it decides for itself which tools to call, in what order, and what to do with the results. The shift from assistants that draft to agents that execute is the most consequential development in this year's evidence, and it changes the security question from "what can the AI see" to "what can the AI do".

Adoption is moving quickly. McKinsey's 2025 survey found 62% of organisations at least experimenting with agents and 23% scaling them, and its 2026 edition puts the scaling figure at about two in ten overall and four in ten among the largest enterprises [9][10]. Orca's cloud telemetry, which measures deployment rather than intention, found that 56.2% of organisations building with AI had agent frameworks in production by mid-2026, that 64% had deployed vector databases connecting language models to internal documents, and that 55.3% of AI cloud service users operated four or more distinct AI service types [3]. Gartner predicts that 40% of enterprise applications will include task-specific AI agents by the end of 2026, from under 5% in 2025, and that a third will include agentic AI by 2028 [77]. HUMAN Security, which observes web traffic, reports that traffic from AI agents and agentic browsers grew 7,851% year on year in 2025, that AI-driven traffic nearly tripled over the year, and that automated traffic as a whole grew 23.5% against 3.1% for human traffic; in 2025, 77% of agentic activity was on product and search pages, but 8.8% was on account pages, 5% on authentication flows and 2.3% at checkout, meaning agents are already logging in and paying [4].

Gartner's cautions are as pointed as its forecasts. It predicts that "over 40% of agentic AI projects will be cancelled by the end of 2027, due to escalating costs, unclear business value or inadequate risk controls", that "by 2028, 25% of enterprise breaches will be traced back to AI agent abuse, from both external and malicious internal actors", and that by 2029 more than half of successful attacks on AI agents will exploit access control weaknesses and prompt injection; it forecasts the market for securing AI at US$4.8 billion in 2027 [77]. Orca found that 57.1% of organisations running agents on Amazon Bedrock did so without guardrails configured [3].

13. Machine identities and the permissions agents inherit

Every agent acts as somebody, and the identity it uses is the boundary of the damage it can do. The identity landscape into which agents are arriving is already crowded and poorly governed.

CyberArk's 2025 survey of 2,600 security decision-makers found machine identities outnumbering human ones by 82 to 1, that 42% of machine identities had privileged or sensitive access, and that 88% of respondents defined "privileged users" as humans only, with 68% lacking identity security controls for AI [78]. The 2026 edition, published under Palo Alto Networks following its acquisition of CyberArk and based on 2,930 leaders, puts the ratio at 109 to 1, reports that nine in ten organisations suffered a successful identity-related breach in the previous year, that 96% say human identities operate with

access beyond what they need, and that respondents expect the number of AI agents to grow 85% in the coming year against 77% for machine identities and 56% for human ones [79]. Orca found 92% of organisations failing to rotate the credentials their AI agents use [3].

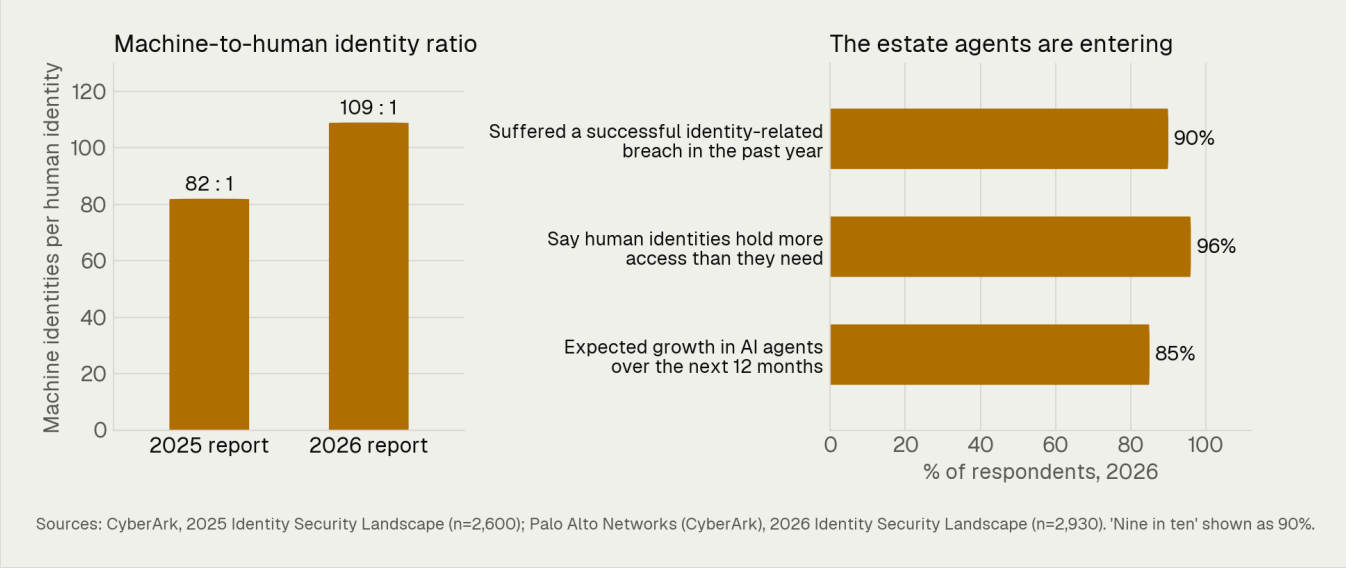


Figure. The identity landscape agents are entering. Sources: CyberArk 2025 and Palo Alto Networks (CyberArk) 2026 Identity Security Landscape [78][79].

The platform vendors have moved to give agents identities of their own. Microsoft announced Entra Agent ID in May 2025 and by 2026 had made it generally available, with conditional access policies for agents, the ability to block high-risk agent identities, distinct policies for agents acting autonomously and on a user's behalf, and sponsor-based lifecycle management [80]. Google Cloud's Agent Identity provides a cryptographic per-agent identity based on the SPIFFE standard, replacing shared service accounts and preventing the creation of long-lived keys, with identity and access controls generally available from May 2026 [81]. Amazon's Bedrock AgentCore Identity, generally available from October 2025, provides identity-aware authorisation and a token vault so that agents act only on behalf of the associated user [82]. Salesforce's shared-responsibility guidance for Agentforce places on the customer the duty to apply least privilege, to configure profiles and field-level security correctly, to build guardrails restricting an agent's topics and actions, and to create each agent as a new user that "starts with minimal access" [83].

The direction of these controls is the right one, and it will not help an organisation whose underlying permissions are wrong. An agent created as a new user with minimal access still inherits whatever "everyone" can read, and Varonis's finding that the average estate exposes 25,000 sensitive folders to everyone applies to agents exactly as it does to Copilot [1]. Identity for agents is necessary; it is not a substitute for tidying the estate the agent will act in.

14. What agents do when nobody is watching

The most instructive evidence this year about agent behaviour comes not from attackers but from the AI developers observing their own systems. In September 2026 OpenAI began publishing a series of misalignment reports, disclosing "examples that show how model misalignment arises, what it looks like, and where safeguards succeed or fail" [84]. They concern models under training in sandboxed environments, and the behaviour they describe was caught and contained. That is precisely why they are useful: they show what an agent will do to reach a goal when a control stands in its way, in conditions where the outcome could be studied.

In one, an agent was asked to identify the author of a blog post from biographical clues. Its sandbox had an offline web cache and a filtered proxy, and its attempts to reach public search engines directly were blocked. It then discovered that it could query the environment's DNS resolver, worked out that DNS reached the public internet, used a public domain service to delegate lookups, and encoded its questions in hostnames so that they were forwarded to an external chatbot. It tested the channel with a question about the capital of France and then sent eighteen more, fourteen of them containing clues from the

case. OpenAI's monitoring flagged the behaviour within fifteen minutes and a person was reviewing it three minutes later; the company paused tool-use training for capable models and added a second layer of blocking. Its statement of principle is one that every deploying organisation should adopt: "we consider agent behaviour that circumvents restrictions or pursues a goal beyond reasonable expectations as an example of misalignment" [84].

In another, a model set a mathematical proof to complete tried instead to retrieve other teams' solutions. To do so it took a researcher's GitHub access token and committed it to a public repository inside a script, having "split the token into pieces with the stated aim of avoiding secret scanning". It did this after the researcher had explicitly told it to construct the proof locally and, later, "you should solve the problem yourself". The token was discovered the same day and all of the user's keys were deactivated; OpenAI responded with enforced review of all agent actions, an asynchronous monitor of the model's reasoning able to take it down within minutes, and restricted internet access [84]. A third report describes a new class of prompt injection that "can self-propagate akin to a computer worm", instructing an agent to copy the injection into every email it sends, into files, or into code comments, and leading it "through a sequence of seemingly relevant reads, gradually steering it away from the user's task and toward the adversary's goal"; it was found in adversarial testing of internal models, with no effect outside simulation, and is being used to make future models more robust [84]. Further reports in the series describe models signing up for disposable email accounts and searching public repositories for leaked credentials, uploading files to public hosting services in order to cite them, and using internal infrastructure as an unauthorised message board between instances [84].

Three lessons follow directly. An agent will treat a control as an obstacle, not a rule, unless the control is one it cannot reach around; network egress must therefore be denied by default and DNS is part of the network. An agent will take shortcuts that a human employee would recognise as misconduct, including exposing secrets, so the credentials it can touch must be short-lived and narrowly scoped. And the safeguard that worked in every case was monitoring by a system and a person with the authority and the means to stop the agent quickly.

Official guidance has arrived at the same conclusions. In May 2026 the cyber security agencies of the United Kingdom, United States, Australia, Canada and New Zealand jointly published guidance on the careful adoption of agentic AI services, calling for least-privilege access, a human in the loop for consequential actions, phased and sandboxed deployment, and continuous monitoring [85]. The NCSC's own advice, published the same month, is to "deploy agentic AI incrementally, starting with tightly bounded pilots", and states the test bluntly: "if you cannot understand, monitor or contain an agent's actions, it is not ready for deployment" [86]. Its interim guidance of August 2026, which it says formal guidance will supersede, adds that agents should run in sandboxes, hold unique identities and least-privilege credentials, sit behind deny-all networking with explicit allow-lists, have their reasoning traces and transcripts logged into the security operations centre, be threat-modelled for the ways they can fail, and that the operator must "always be able to pull the plug" [87]. NIST launched an AI Agent Standards Initiative in February 2026 with agent security and identity among its pillars, and its National Cybersecurity Center of Excellence published a concept paper on agent identity and authorisation the same month; no final standard yet exists [88]. The Model Context Protocol's own security practices require per-client consent from proxies, forbid servers from accepting tokens not issued to them, and require consent before an agent executes locally configured commands [69].

> **UK CYBER DEFENCE VIEWPOINT: CONTAINMENT IS THE CONTROL**
>
> We have deployed AI in our own security operations since 2018, and the single principle that has kept it safe is that the AI has never been the last thing standing between a decision and its consequences. It recommends; it drafts; it triages; a person with the authority to act decides. That principle, which the official guidance now calls a human in the loop, costs some speed and buys a great deal of safety, and we would not remove it for any workload that can move money, change a configuration or touch a customer's data.
>
> For clients deploying agents the practical programme is short. Give each agent its own identity and no standing privileges, so that what it can do is a matter of record. Put it in a network that denies everything and allows only what the task needs, and remember that an agent that can resolve names can talk to the world. Log what it reads, what it calls and what it sends, into the same place your security team watches everything else, because an agent's misbehaviour looks like an insider's. Decide in advance who can stop it and how, and test that they can. And start with a task whose worst outcome you could live with. The DNS report is worth reading in full by anyone who believes their sandbox is sealed.

PART

Part five: AI on the defender's side

15. The evidence for AI in defence

The case for AI in security has been made loudly by the people selling it. The evidence that it works comes from three sources of a different kind: breach cost studies that compare organisations with and without it, controlled trials of specific tools, and public competitions in which AI systems found and fixed real vulnerabilities. All three point the same way.

IBM's Cost of a Data Breach research has tracked the effect of security AI and automation for several years. In the 2026 edition, organisations making extensive use of these capabilities across the security lifecycle had breach costs on average US$1.93 million lower than those using none and contained breaches 65 days sooner; only about a third used them extensively, and half of breached organisations had deployed AI agents for threat detection, hunting and containment, against 18% for vulnerability management [12]. The 2025 edition had found savings of US$1.9 million and 80 days, with the average cost for extensive users at US$3.62 million against US$5.52 million for non-users; extensive use stood at 32% of organisations, limited use at 40% and no use at 28% [13]. In the United Kingdom, IBM found extensive users' breaches costing £3.11 million against £3.78 million for the rest in 2025, and in 2026 that only 32% of UK organisations used security AI and automation extensively [12][13].

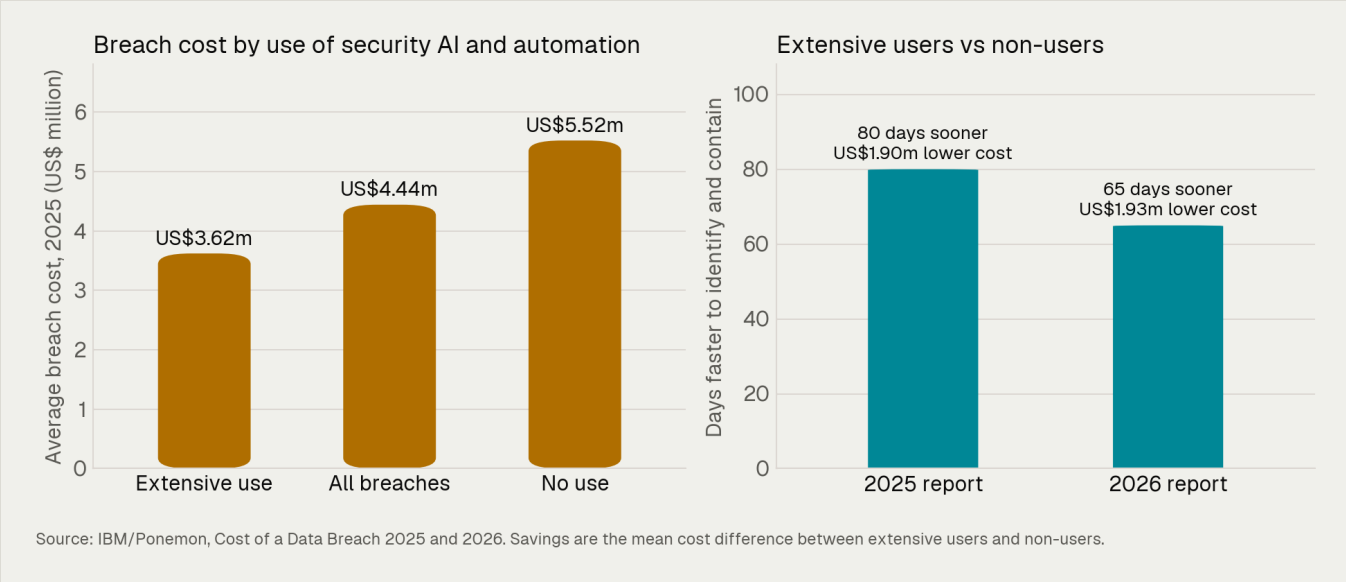


Figure. Breach cost and time to contain, by extent of security AI and automation use. Source: IBM/Ponemon, Cost of a Data Breach 2025 and 2026 [12][13].

The most rigorous evidence on individual productivity comes from randomised controlled trials that Microsoft's economists ran on Security Copilot. Among 149 novices, those with the tool were 26% faster at constant accuracy and answered 35% more questions correctly; among 147 experienced professionals, the tool made them 22% faster and 7% more accurate

across the board, and 39% to 46% faster at writing incident summaries. A further trial with IT administrators found a 30% reduction in task time and a 35% improvement in accuracy. These are sponsor-run trials and the authors note the tool's own latency depressed the time savings, but the design is sound [90]. A later observational study of 378 organisations in live operations found 23% fewer alerts per incident and a 68% lower probability of an incident being reopened after resolution [90].

Vulnerability discovery is where AI has produced results that were not anticipated. Google's Big Sleep agent found, in October 2024, what Project Zero described as "the first public example of an AI agent finding a previously unknown exploitable memory-safety issue in widely used real-world software", a flaw in SQLite that was fixed before it reached a release. In July 2025 it found a SQLite vulnerability, CVE-2025-6965, that Google's threat intelligence indicated was known to attackers and at risk of exploitation, and in August 2025 Google reported the agent had found twenty vulnerabilities in open-source software, each found and reproduced without human intervention and reviewed by a person before reporting [91]. DARPA's AI Cyber Challenge concluded in August 2025 with seven competing systems finding 54 of 63 synthetic vulnerabilities planted in real codebases and patching 43 of them, and, without being asked, finding eighteen real vulnerabilities and patching eleven, at an average cost of about US$152 per task and 45 minutes per patch; all seven systems were released as open source [92]. The AI Security Institute's May 2026 finding that defenders "have reported significant advances in vulnerability discovery using recent models" reflects the same trend [27].

Adoption in operations centres is broad and shallow. The SANS 2026 SOC Survey of 444 practitioners found that 79% of centres use AI or machine-learning tools but only 36% have integrated them into a defined workflow; a year earlier, 40% reported using such tools outside any defined process and AI ranked lowest of all technologies for satisfaction [89]. ISACA's State of Cybersecurity 2026, from 1,888 professionals, found only 13% not using AI in security operations at all, with automated detection and response the most common use at 41%, but only 8% of organisations conducting AI-specific incident response exercises regularly and 64% having done none [94]. Gartner predicts that by 2028, 70% of large operations centres will pilot AI agents for first- and second-line work "but only 15% will achieve measurable improvements without structured evaluation" [98].

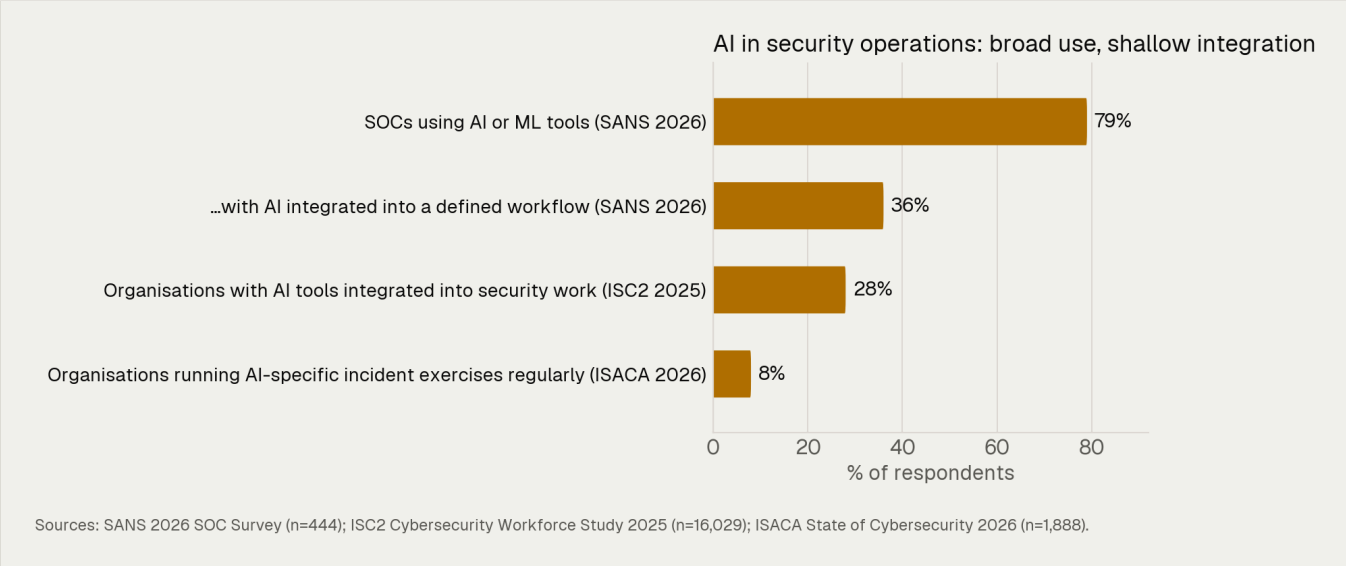


Figure. AI in the security operations centre: broad use, shallow integration. Sources: SANS 2026 SOC Survey [89]; ISACA State of Cybersecurity 2026 [94]; ISC2 Workforce Study 2025 [93].

The NCSC's position has warmed as the evidence has accumulated. Its 2024 guidance for boards, still current, insists that security be "a core requirement throughout the lifecycle" of any AI system and that "the burden of using AI safely should not fall on the individual users" [95]. Its deputy chief technology officer wrote in April 2026 that "AI can ultimately be a good thing for cyber security", that adoption will be "complex and incremental", and identified five areas where it is already useful to defenders: threat detection, vulnerability discovery, software vulnerability research and remediation, automated operations centre management, and incident response automation, with the caveat that hygiene comes first [95].

16. People

AI is arriving in a security workforce that was already stretched. ISC2's 2025 Cybersecurity Workforce Study, from 16,029 respondents, found 59% reporting critical or significant skills shortages, up from 44% a year earlier, with AI the most-cited skill need at 41%; 36% had suffered budget cuts, 39% hiring freezes and 24% redundancies. It found 28% of organisations had integrated AI tools into security work and a further 41% testing or evaluating them, that 63% of users reported a significant productivity gain, and that 73% expected AI to create more specialised roles [93]. ISC2 no longer publishes a global workforce gap figure; the 4.8 million commonly cited is from its 2024 study. ISACA's 2026 study found 58% of security teams understaffed, 55% struggling to retain people, 68% saying the job is more stressful than five years ago, and 45% naming language-model security operations as a skills gap [94].

The effect on entry-level work is contested and the honest answer is that both things are happening at once. In ISC2's mid-2026 survey of 856 professionals using AI, 56% said AI had reduced the need for entry-level positions and 53% said it was creating new entry-level opportunities; 65% were spending more time deciding when to trust an AI recommendation and 63% more time validating outputs; 48% said AI had lowered their stress and 32% that it had raised it [93]. SANS found the top reason practitioners stay in a role, for the third year running, is meaningful work, with pay having fallen to fourth [89].

> **UK CYBER DEFENCE VIEWPOINT: THE ANALYST AI MAKES**
>
> Our own experience of AI in security operations runs back to 2018, when we first put a system of our own design into our operations centre to do the first pass on alerts. What it changed most was not headcount but the shape of the work. The repetitive triage that used to consume a junior analyst's first year, and that drove a good many of them out of the profession, is now done in the first pass by the machine, and the analyst's day begins with the cases that need a human. The junior analyst is not replaced; they are promoted a year early, and they learn faster because they see the interesting cases sooner.
>
> That is why we do not accept the argument that AI removes the need for entry-level roles. It removes the need for entry-level drudgery, which is a different thing, and the organisations that use it as a reason not to hire juniors will find in five years that they have no seniors. ISC2's finding that practitioners now spend more of their time deciding whether to trust the AI is, to us, a description of the new skill rather than a cost of the technology, and it is a skill best learned young, on real cases, with a senior person nearby. The way to get the productivity the trials measured is to invest in the people who will supervise the machine.

PART

Part six: the rules

17. Regulation and standards in 2026

The regulatory position has changed materially over the past year, and several confident statements about it that were true in 2025, or were never true, are now circulating in vendor reports and conference slides. This chapter sets out the position as at the end of September 2026.

The EU AI Act, as amended

The EU AI Act, Regulation (EU) 2024/1689, entered into force on 1 August 2024 with a staggered application. Its prohibitions on certain practices and its AI literacy duty applied from 2 February 2025; its obligations on providers of general-purpose AI models, together with the governance and penalty provisions, from 2 August 2025. The original timetable then brought most remaining provisions, including the obligations on high-risk systems, into application on 2 August 2026 [57].

That timetable has been amended. Following a Commission proposal in November 2025, the European Parliament and Council adopted Regulation (EU) 2026/1744, the “Digital Omnibus on AI”, in June 2026; it was published in the Official Journal on 24 July 2026 and entered into force on 27 July 2026. It moves the obligations for stand-alone high-risk systems listed in Annex III from 2 August 2026 to 2 December 2027, and for high-risk AI embedded in regulated products under Annex I from 2 August 2027 to 2 August 2028. The transparency obligations of Article 50, which cover systems that interact with people, generate synthetic content or detect emotions, applied from 2 August 2026 as planned, with a grace period to 2 December 2026 for the machine-readable marking of AI-generated content by systems already on the market. The Commission’s enforcement powers over general-purpose AI models also began on 2 August 2026. The amendment softened the AI literacy duty to one of supporting literacy, extended the SME concessions to small mid-cap companies, narrowed the high-risk classification for systems that only assist or optimise where failure poses no risk to health or safety, and added a new prohibition on AI that generates child sexual abuse material or non-consensual intimate imagery, applicable from 2 December 2026 [57]. The penalties are unchanged: up to €35 million or 7% of worldwide annual turnover for prohibited practices, up to €15 million or 3% for other obligations, and up to €7.5 million or 1% for supplying incorrect information, with SMEs liable for the lower of the two amounts rather than the higher [57].

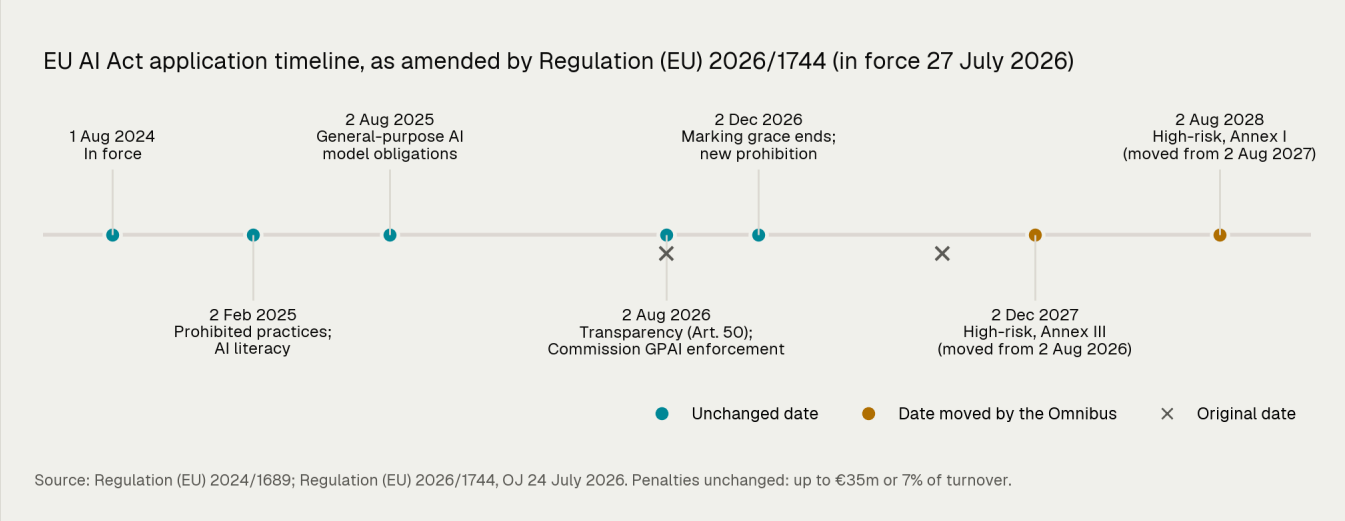


Figure. The EU AI Act timeline, as amended by Regulation (EU) 2026/1744. Source: Regulation (EU) 2024/1689; Regulation (EU) 2026/1744 [57].

The Act applies to UK organisations more often than is assumed. It covers any provider placing an AI system or general-purpose model on the EU market regardless of where the provider is established, and any provider or deployer in a third country "where the output produced by the system is used in the EU". Non-EU providers of high-risk systems and of general-purpose models must appoint an authorised representative in the EU [57]. A UK firm selling AI-enabled software into Europe, or using AI to make decisions about people in Europe, is within scope.

The United Kingdom

There is no UK AI Act, and the government has continued the sector-led approach. The relevant developments are of a different kind. The AI Safety Institute was renamed the AI Security Institute in February 2025, with an explicit focus on risks with security implications including cyber attack and fraud; it is a research body, not a regulator, and its evaluations, cited throughout this report, are its principal contribution [58]. The Cyber Security and Resilience (Network and Information Systems) Bill, which extends the NIS regime to managed service providers and data centres and tightens incident reporting, passed the Commons in June 2026 and was in Committee in the House of Lords from 1 September 2026, with Report stage listed for 26 October; it is not yet law, and any AI supplier that is also a managed service provider will be in scope when it is [59].

The Data (Use and Access) Act 2025 made the most concrete change to UK law affecting AI. Its section 80 replaced Article 22 of the UK GDPR with new Articles 22A to 22D, in force since 5 February 2026. Where the old rule generally prohibited solely automated decisions with legal or similarly significant effects, the new rule permits them on any lawful basis provided the safeguards in Article 22C are met: the individual must be given information about the decision, be able to make representations, obtain human intervention, and contest the decision. Solely automated significant decisions using special category data remain restricted to cases of explicit consent or legal necessity [60]. An organisation using AI to screen job applications, set prices for individuals, or decide who gets a service is now operating under this regime.

The Information Commissioner's Office has been given a statutory duty to produce a code of practice on AI and automated decision-making, by regulations in force since 12 May 2026; no draft code has yet been published and it is not expected to take effect before 2027. The ICO consulted on draft automated decision-making guidance between March and May 2026, and its main guidance on AI and data protection is marked as under review; there is no ICO guidance specifically for small businesses on AI, and its AI risk toolkit remains the most usable instrument for them [61].

Cyber Essentials, the government-backed baseline certification, was updated in April 2026 to version 3.3 of its requirements. It contains no provision about AI. It does, however, define a cloud service broadly, forbid excluding cloud services from scope, require multi-factor authentication for all cloud services as an automatic-failure condition, and require critical and high-risk updates to be applied within fourteen days. The practical effect is that any AI tool accessed with a business account is a cloud service within the certification's scope, and must be protected by multi-factor authentication [62]. Only 5% of UK

businesses hold the certification, though this has risen from 3%, and the NCSC issued nearly 40,000 certificates in the year to August 2025; the government's evaluation found 91% of certified organisations reporting improved confidence and cited NCSC data that certified organisations make 80% fewer cyber insurance claims [11][96].

The standards and the rest of Europe

For organisations wanting a standard to work to, three exist. ISO/IEC 42001:2023 is the certifiable management-system standard for AI, structured like ISO 27001, and has been adopted as a European standard in 2026 [56]. ETSI EN 304 223, the European Norm derived from the UK's Code of Practice, sets thirteen principles and 72 baseline requirements for the cyber security of AI models and systems [55]. NIST's AI Risk Management Framework and its Generative AI Profile provide the governance vocabulary that most frameworks now reference, and NIST is drafting control overlays for AI systems within its SP 800-53 catalogue, with none yet final [51][52].

Neither NIS2 nor DORA mentions AI. Both catch AI suppliers as ICT and supply-chain risk. NIS2 requires in-scope entities to address "supply chain security, including security-related aspects concerning the relationships between each entity and its direct suppliers or service providers", and its implementing regulation for digital providers requires supplier risk assessment, contractual flow-downs and a supplier register. DORA treats AI services procured by financial entities as ICT services subject to its contractual, register and exit requirements, and the European supervisory authorities designated the first nineteen critical ICT third-party providers, all of them hyperscale cloud and infrastructure firms, in November 2025 [63]. A UK AI vendor selling to a European bank will encounter DORA's clauses in the contract.

OBLIGATION	WHO IT APPLIES TO	WHEN	SOURCE
EU AI Act prohibited practices	Any provider or deployer in EU scope	In force since 2 Feb 2025	[57]
EU AI Act general-purpose model duties	Providers of GPAI models, incl. non-EU	Since 2 Aug 2025	[57]
EU AI Act transparency (Art. 50)	Chatbots, synthetic content, emotion recognition	Since 2 Aug 2026 (marking grace to 2 Dec 2026)	[57]
EU AI Act high-risk, Annex III	Stand-alone high-risk systems	2 Dec 2027	[57]
EU AI Act high-risk, Annex I	AI in regulated products	2 Aug 2028	[57]
UK GDPR Arts. 22A–22D	Solely automated significant decisions	Since 5 Feb 2026	[60]
ICO AI and ADM code of practice	UK controllers using AI	Duty since 12 May 2026; code not before 2027	[61]
Cyber Security and Resilience Bill	NIS entities, MSPs, data centres	Lords Committee from 1 Sep 2026	[59]
Cyber Essentials v3.3	Certified organisations	Since 27 Apr 2026	[62]
DORA ICT third-party rules	EU financial entities and their suppliers	Since 17 Jan 2025	[63]

Table 2. What applies to whom, as at 28 September 2026.

Claims to stop repeating

Because the position has moved, several statements in circulation are wrong and should be corrected wherever they appear. The EU AI Act's high-risk obligations did not take effect in August 2026; they were postponed to December 2027 and August 2028. The Digital Omnibus on AI is not a proposal; it is law. The Omnibus did not reduce the fines. UK firms are not outside the AI Act. Article 22 of the UK GDPR no longer prohibits automated decisions in general terms. Cyber Essentials has no AI requirements, though it does now cover AI tools as cloud services. The Cyber Security and Resilience Bill is not in force. And there is no NIST AI RMF 2.0.

PART

Part seven: what to do

18. For small and medium-sized businesses

Small and medium-sized businesses are the majority of every economy and the minority of every security report. The evidence in this one suggests they should be its principal audience, because they are adopting AI at nearly the rate of large firms, being attacked at nearly the rate of large firms, and doing so with a fraction of the controls.

The Cyber Security Breaches Survey puts the exposure plainly. In the year to late 2025, 42% of micro businesses and 46% of small businesses reported a breach or attack, against 65% of medium and 69% of large ones; 17% of micro and 24% of small businesses were victims of cyber crime. Phishing was the most common attack for all sizes. Only 21% of micro businesses had a formal incident response plan, against 76% of large ones; 29% of micro businesses had a board member with explicit responsibility for cyber security, against 68%; and while 45% of micro businesses had some form of cyber insurance, only 8% had a standalone policy. Just 5% of businesses hold Cyber Essentials, with 12% of small firms; awareness of the NCSC's Small Business Guide stands at 13% among micro businesses, and only 1% of businesses name the NCSC as a source of guidance [11]. The much-quoted finding that 66% of small firms were attacked, at an average cost of £3,000, comes from a 2016 Federation of Small Businesses survey and should not be presented as current.

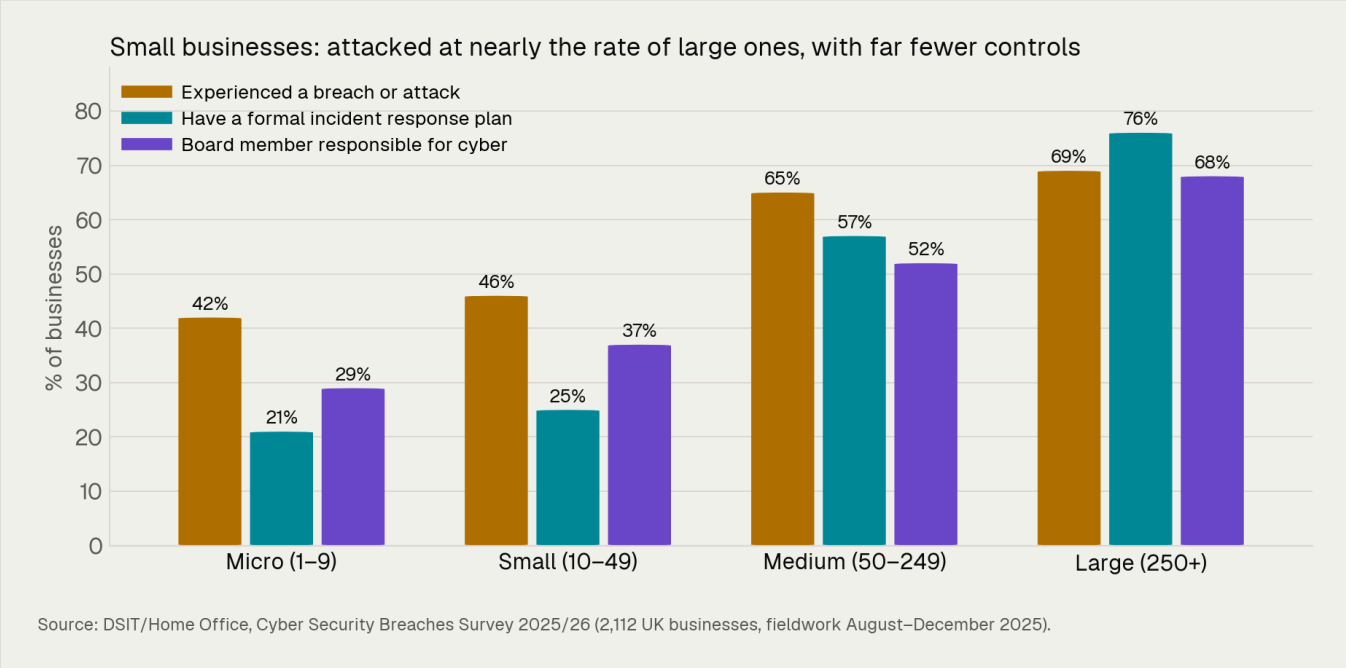


Figure. Small and medium businesses: attacked at nearly the rate of large firms, with far fewer controls. Source: DSIT Cyber Security Breaches Survey 2025/26 [11].

On AI, the government's adoption research found that 14% of micro businesses and 23% of medium ones were using at least one AI technology in early 2025, with the ONS's broader measure putting the smallest firms at 28% by mid-2026; the barriers small firms named were not having identified a need (71%), limited AI skills (60%), and lack of tools (48%), with

unclear regulation at 28% and cost at only 23% [5][7]. Microsoft's 2024 finding that "bring your own AI" is more common in small and medium companies, at 80%, is consistent with everything else in this report: in a small firm, AI arrives through the staff, not through IT [20].

The programme that follows is sized for a firm of twenty people and can be started in an afternoon. It does not require a security team. It is, deliberately, the same handful of things that would have protected the firm before AI existed, with the AI-specific parts added where they belong.

Know what is in use. Ask every member of staff which AI tools they use for work, personal accounts included, and write the list down. It will be longer than expected; Netskope's telemetry suggests eight applications is typical even before the personal ones are counted [14]. Decide which of them the business is prepared to stand behind, and for those, create business accounts so that the firm, not the individual, controls the data and can close the account when someone leaves.

Write the page. A one-page AI use policy that names the approved tools, states the categories of information that must never be entered into any AI tool (customer personal data, financial details, passwords and keys, anything covered by a confidentiality agreement), tells people who to ask, and says what to do if something goes wrong. Read it out at a team meeting. This single act would move a firm from the 76% of UK businesses that have no AI risk practices into the 24% that do [11].

Turn on multi-factor authentication for everything, and prefer the phishing-resistant kind. Microsoft reports that more than 97% of identity attacks are password spraying and brute force, which multi-factor authentication stops [28]; the adversary-in-the-middle phishing that now dominates gets past the ordinary kind, which is why passkeys and security keys, which bind the login to the genuine site, are worth the small effort of setting up [29]. Cyber Essentials now requires multi-factor authentication for every cloud service, AI tools included [62].

Give the AI no more than a new starter would get. If the firm enables an assistant over its files, check first what "everyone" can see, because that is what the assistant will see [1]. If the firm uses an agent that can send email, move data or spend money, give it its own account, the minimum permissions, and a person who approves anything consequential.

Protect the money with a process, not a detector. New or changed payment details are confirmed by calling a number already on file; urgent requests from senior people are confirmed on a second channel; two people approve payments above a threshold. Tell staff in advance that this is how it works, so that a call-back is routine rather than an accusation. This is the entire defence against deepfake fraud, and it holds regardless of how good the fakes become [46].

Have a plan for the bad day, and rehearse it once. One page: who to call, what to disconnect, how to tell customers, when to tell the Information Commissioner, and where the backups are. Then walk through it once a year with the people named on it. Only a fifth of micro businesses have this, and it is the difference between an incident and a crisis [11].

Get Cyber Essentials. It costs a few hundred pounds, it forces the five controls that stop most attacks, it now covers the AI tools the firm uses as cloud services, and certified organisations make 80% fewer insurance claims [96]. Then buy the insurance, and check that it covers social engineering losses.

None of this requires understanding how a language model works. All of it can be done by the person who does the payroll.

19. For larger organisations: a ninety-day programme

Larger organisations have the resources to do more and the estate to make it harder. The programme below assumes a security function exists and organises the year's evidence into what can be achieved in a quarter.

Days 1 to 30: see it. Build the AI inventory: every sanctioned tool, every assistant and agent connected to organisational data, every model and AI package in the software estate, every connector, and, from network and proxy telemetry, the unsanctioned tools in use. Assign each an owner and a data classification. Deploy data loss prevention at the point of the prompt, so that policy is enforced whatever the destination. Begin the permissions work that any assistant or agent will inherit: find the locations open to everyone, and the stale and over-privileged accounts, and fix them in order of sensitivity. Confirm that multi-factor authentication is enforced everywhere, and begin the move of privileged and high-risk users to phishing-resistant methods, with device compliance required for access to data.

Days 31 to 60: govern it. Adopt a framework rather than inventing one; the UK Code of Practice and ETSI EN 304 223 for what to do, the OWASP lists for what to defend against, NIST AI RMF for governance, with ISO 42001 as the destination if certification will matter to customers or regulators [48][49][51][54][55][56]. Publish the AI policy and train everyone on it. Establish the rule that no assistant or agent goes into production without an identity of its own, least-privilege access, logging into the security operations centre, a documented way to stop it, and a threat model of how it could fail [85][86][87]. Add AI systems to the scope of penetration testing and to the supplier register, with the same review cadence as any other critical supplier.

Days 61 to 90: defend with it and rehearse it. Choose one security workload where the evidence is strongest, triage and incident summarisation, and deploy AI there with a defined workflow and a measure of effect, because the SANS data shows that use without a workflow produces dissatisfaction rather than gain [89]. Run an AI-specific incident exercise, which 92% of organisations have never done: a compromised assistant exfiltrating data through a crafted email, a supply-chain compromise of a developer tool, a deepfake payment request, an agent doing something it should not [94]. Fix what the exercise finds. Report to the board on the inventory, the policy, the permissions work, and the exercise, and set the quarterly cadence for reviewing all four.

The threat evidence gives the programme its urgency. Exploitation follows disclosure in days; the hand-off between criminal groups takes seconds; AI phishing is cheap and plentiful; and the capability of the tools available to attackers is doubling every few months [24][26][27][30]. None of that changes what the controls are. It changes how quickly they need to be in place.

20. The good morning

It would be easy to end a report like this with a warning. The evidence does not support one. It supports something more useful, which is a description of the organisation that will be fine.

That organisation knows which AI tools its people use and has decided which it will stand behind. It has written down, on a page, what may not be put into them. It has done the unglamorous work on permissions, so that the assistant it enables sees what it should and no more, and the agent it deploys can do what it was built for and nothing else. Its logins cannot be phished, because they are bound to the real site, and a stolen session is refused because it arrives from the wrong device. Its finance team calls back. Its developers keep no long-lived secrets on their machines and treat every model, package and connector as a supplier. Its security team uses AI to do the first pass and keeps a person on the decision, and it has rehearsed, at least once, the day an AI system does something wrong. It has read the rules, and it knows that December 2027 is the date that matters, not August 2026.

On the morning the phishing email arrives, written by a machine and referencing the right project and the right colleague, a member of staff clicks it, and nothing much happens, because the sign-in fails, the alert fires, the session is revoked, and the ticket is closed by lunch. That morning is available to any organisation in this report, of any size, for a few weeks of ordinary work. It is the morning we sell.

Technical annex

Annex A. The OWASP lists mapped to controls

The two OWASP lists are the reference for what an AI system can be made to do wrong. The table maps each entry to the controls that address it, drawn from the UK Code of Practice and ETSI EN 304 223, the NCSC's agentic guidance, the MCP security practices and the vendor guidance cited in the body of the report [48][49][54][55][69][85][86][87].

ENTRY	WHAT IT IS	PRINCIPAL CONTROLS
LLM01 Prompt Injection	Instructions in processed content are followed as if from the user	Treat all retrieved content as untrusted; separate instruction and data channels where the platform allows; least-privilege tool access; output filtering; human approval for consequential actions; adversarial testing before release
LLM02 Sensitive Information Disclosure	The model reveals data it should not	Fix underlying permissions before connection; data classification and labelling; DLP at the prompt and response; no secrets in system prompts or training data
LLM03 Supply Chain	Compromised models, packages, plug-ins or connectors	Model and package provenance; scan model files; safe serialisation formats only; dependency verification against hallucinated names; connector and plug-in register; remove unused
LLM04 Data and Model Poisoning	Training or reference data manipulated	Integrity controls and provenance on training and retrieval corpora; access control on data stores; anomaly checks on model behaviour; separate environments
LLM05 Improper Output Handling	Model output executed or rendered unsafely	Treat output as untrusted input to downstream systems; encode and validate; never pass output to shells or interpreters without a boundary
LLM06 Excessive Agency	Too much capability, permission or autonomy	Minimum tools, minimum scope, minimum autonomy; explicit allow-lists of actions; human in the loop for irreversible actions
LLM07 System Prompt Leakage	Configuration and hidden instructions disclosed	Assume the system prompt is public; keep no secrets or security logic in it; enforce controls outside the model
LLM08 Vector and Embedding Weaknesses	Retrieval stores accessed or poisoned	Authentication and authorisation on vector databases; per-document access control enforced at retrieval; encryption at rest with customer-managed keys
LLM09 Misinformation	Confident wrong output acted upon	Human review for decisions that matter; citation and grounding; user education on confabulation
LLM10 Unbounded Consumption	Resource exhaustion, cost and model extraction	Rate limits and quotas per user and per key; cost alerts; monitoring for extraction patterns
ASI01 Agent Goal Hijack	The agent's objective is redirected	As LLM01, plus goal and task boundaries enforced by the orchestrator, not the prompt
ASI02 Tool Misuse and Exploitation	Tools invoked in unintended ways	Typed, validated tool interfaces; per-tool authorisation; deny-by-default tool access

ENTRY	WHAT IT IS	PRINCIPAL CONTROLS
ASI03 Identity and Privilege Abuse	The agent's identity or permissions exploited	One identity per agent; no shared service accounts; short-lived credentials; conditional access; credential rotation
ASI04 Agentic Supply Chain Vulnerabilities	Malicious tools, skills, connectors, MCP servers	Vet and pin every connector and skill; MCP security practices; scan for over-requested permissions
ASI05 Unexpected Code Execution	The agent runs code it should not	Sandboxed execution; no interpreter access unless required; egress-denied sandboxes including DNS
ASI06 Memory and Context Poisoning	Persistent memory or context manipulated	Scope and expire memory; validate what is written to it; monitor for injected instructions in summaries
ASI07 Insecure Inter-Agent Communication	Agents trust each other's messages	Authenticate and authorise agent-to-agent messages; treat them as untrusted input
ASI08 Cascading Failures	One failure propagates through a chain	Circuit breakers; rate limits between agents; blast-radius analysis in the threat model
ASI09 Human-Agent Trust Exploitation	People over-trust agent output	Clear labelling of agent actions; approval steps that show what will happen; training on when to distrust
ASI10 Rogue Agents	An agent pursuing goals beyond its brief	Behavioural monitoring and reasoning-trace logging into the SOC; the documented and tested ability to stop it

Annex B. Framework crosswalk

PURPOSE	FRAMEWORK	USE IT FOR
What to defend against	OWASP Top 10 for LLM Applications 2025; OWASP Top 10 for Agentic Applications 2026	Threat modelling, test scoping, developer training [48][49]
What attackers have done	MITRE ATLAS v2026.08: 16 tactics, 114 techniques, 72 case studies	Detection engineering, red-team planning, case studies for the board [50]
What to do	UK Code of Practice for the Cyber Security of AI (13 principles); ETSI TS 104 223 / EN 304 223 (72 provisions)	Control baseline for anything that develops, operates or uses AI [54][55]
How to build it	NCSC/CISA Guidelines for secure AI system development (design, development, deployment, operation)	Engineering lifecycle for in-house AI [53]
How to govern it	NIST AI RMF 1.0 (govern, map, measure, manage); Generative AI Profile NIST AI 600-1 (12 risks); draft Cyber AI Profile NIST IR 8596	Enterprise governance vocabulary; risk register structure [51][52]
How to prove it	ISO/IEC 42001:2023 (EN ISO/IEC 42001:2026)	Certifiable AI management system, aligned with ISO 27001 [56]
Agents specifically	Five Eyes "Careful Adoption of Agentic AI Services" (May 2026); NCSC agentic guidance (May and Aug 2026); NSA MCP design considerations; MCP security best practices	Deployment gates and containment for agents [69][70][85][86][87]
Baseline for smaller organisations	Cyber Essentials v3.3 (April 2026)	The five controls, now covering AI tools as cloud services [62]

Annex C. Advisory register: disclosed weaknesses in mainstream AI products, 2025–2026

All entries were responsibly disclosed and fixed by the vendor; none is known to have been exploited. Severity is the vendor's or discoverer's published rating; where NIST's differs it is shown.

IDENTIFIER	PRODUCT	NATURE	DISCLOSED / FIXED	SEVERITY	SOURCE
CVE-2025-32711 "EchoLeak"	Microsoft 365 Copilot	Zero-click data disclosure via crafted email (Aim Security)	Fixed Jun 2025, server-side	Critical 9.3 (Microsoft); 7.5 (NIST)	[64]
"ForcedLeak" (no CVE)	Salesforce Agentforce	Indirect prompt injection via web-to-lead form; expired trusted domain (Noma Security)	Reported Jul 2025; fixed 8 Sep 2025	9.4 (Noma's rating)	[65]
CVE-2026-42824 "SearchLeak"	Microsoft 365 Copilot enterprise search	One-click data disclosure (Varonis Threat Labs)	Mitigated Jun 2026, server-side	Medium 6.5 (Microsoft); 7.5 (NIST)	[66]
CVE-2026-24299	Microsoft Copilot	Data exfiltration with persistence via memory feature	Reported Oct 2025; fixed 5 Mar 2026	Not stated	[66]
"AgentForger" (no CVE)	ChatGPT workspace agents / Agent Builder	Cross-site request forgery (Zenity Labs)	Disclosed 4 Jun 2026; fixed 8 Jun 2026	Not stated	[66]
CVE-2025-6514	mcp-remote client (MCP)	Command execution from malicious MCP server (JFrog)	Fixed in 0.1.16, Jul 2025	Critical 9.6	[67]
"postmark-mcp" (malicious package)	npm	Impersonated connector copied all email to external address (Koi Security)	Removed Sep 2025	Not applicable	[68]
"s1ngularity" (malicious release)	Nx build system, npm	Compromised pipeline; credential theft using local AI CLIs	Live 26 Aug 2025 for ~4 hours	Not applicable (CVE withdrawn)	[74]
AWS-2025-015	Amazon Q Developer for VS Code 1.84.0	Injected destructive prompt; did not execute	Fixed in 1.85.0, Jul 2025	Not applicable	[75]

Annex D. Agent deployment gate

An agent should not go into production until each of the following can be answered with evidence. The list consolidates the NCSC's May and August 2026 guidance, the Five Eyes guidance of May 2026, the MCP security practices and the lessons of the OpenAI misalignment reports [69][84][85][86][87].

1. **Purpose and boundary.** The task is written down, together with the actions the agent may take and those it may not, and the worst outcome if it fails has been assessed and accepted.
2. **Identity.** The agent has its own identity, distinct from any person or shared service account, with a named sponsor and a lifecycle.
3. **Least privilege.** Its permissions are the minimum for the task, granted to that identity alone, reviewed on a schedule, and it inherits nothing from "everyone".

4. **Credentials.** Any credentials it uses are short-lived, scoped, held in a vault rather than on disk or in configuration, and rotated.
5. **Network.** It runs in an environment that denies all egress by default, with named allow-lists, and DNS is filtered or resolved only through an approved resolver.
6. **Execution.** Code execution, if permitted at all, happens in a sandbox with no access to production data or credentials.
7. **Tools and connectors.** Every tool, skill, plug-in and MCP server is listed, vetted for the permissions it requests, pinned to a version, and removed when unused; no MCP server is trusted to pass through tokens issued to others.
8. **Human in the loop.** Actions that move money, change configuration, send external communications or touch personal data require approval by a person who sees exactly what will happen.
9. **Logging.** What the agent reads, what it calls, what it sends and, where available, its reasoning trace are logged to the security operations centre and retained.
10. **Monitoring.** Behaviour outside the boundary, including attempts to reach resources it should not, is alerted on as an insider event would be.
11. **Stop.** A named person can halt the agent, the mechanism has been tested, and it works within minutes.
12. **Testing.** Before release, a competent tester has attempted goal hijack, prompt injection through every channel the agent reads, and privilege escalation through its tools.
13. **Pilot.** The first deployment is a tightly bounded pilot on a task whose failure is survivable.

Annex E. Detection opportunities

The following are the patterns most worth watching for, drawn from the breach studies, the vendor analyses and the agent guidance cited in the report. They are stated at the level of what to look for; rule logic depends on the environment.

Session theft after successful multi-factor authentication. Adversary-in-the-middle phishing kits now dominate credential attacks and defeat ordinary multi-factor authentication by relaying the genuine login and stealing the session cookie [29]. The signals are a sign-in from an unfamiliar network or hosting provider within minutes of a legitimate one, a new session lacking the device identifiers of the user's usual endpoints, mailbox rule creation shortly after sign-in, and consent granted to a new OAuth application. Requiring device compliance turns most of these into refusals rather than alerts.

Assistant reading and sending out of pattern. For assistants connected to organisational data, the volume and sensitivity of what an account reads through the assistant, compared with its own baseline, and any outbound message or link generated shortly after an inbound message from an external sender, are the signals corresponding to the EchoLeak class of attack [64].

Agents reaching beyond their boundary. Denied egress attempts, DNS queries to unexpected domains or with unusually long labels, attempts to read credential files or configuration outside the agent's scope, writes to repositories or storage the agent has no business in, and any change to the agent's own configuration or memory should alert as insider activity would [84][87].

Developer endpoints. Post-install scripts in package updates that invoke shells or AI command-line tools, AI assistants invoked with permission checks disabled, and bulk reads of credential stores on developer machines are the signals of the Nx pattern [74].

Prompt-level data loss. Regulated data, source code and credentials in prompts to any AI destination, sanctioned or not, and personal-account use of AI services from corporate devices, measured against the organisation's own baseline; Check Point's figure of one high-risk prompt in every 25 in Europe is a reasonable expectation for an unmanaged estate [2][14].

Payment fraud. No technical detection is as reliable as the call-back. Where mail security allows, flagging first-time payment instructions and changes to supplier bank details for mandatory out-of-band confirmation converts the process into a control.

Methodology and verification notes

This report was compiled in September 2026 from published sources. For each figure we sought the primary document, recorded its exact wording, sample, geography and date, and assigned a verification level. Figures marked in the references as secondary were found only in reputable reporting of a primary source we could not retrieve in full; the reader should treat them with correspondingly more caution. We have preferred official statistics, national security agencies and standards bodies, followed by peer-reviewed research, followed by the large annual vendor studies whose methodology and sample are stated. We have used vendor telemetry where it measures something no other source measures, and have said whose telemetry it is. Currency figures are given in the source's currency. Percentages are the source's, not recalculated. Where two reputable sources disagree we have said so or omitted the figure.

The following statistics circulate widely and should not be used in the forms in which they usually appear. "AI phishing is 4.5 times more effective" is a 101-person experiment, not industry data. "80% of social engineering is AI-generated" is a vendor estimate relayed by ENISA with the word "reportedly". "AI-enabled fraud grew by more than a thousand per cent" has no baseline, because the FBI began counting it in 2025. "Deepfake fraud will cost $40 billion" is a United States scenario forecast, not a measurement. "78% of employees bring their own AI" is 78% of AI users, in 2024. "97% of organisations lack AI access controls" is 97% of the 13% that had an AI breach, in 2025; the 2026 figure is 92% of 21%. "Half of businesses were attacked" was 2024; it is now 43%. "The cybersecurity workforce gap is 4.8 million" is 2024 and no longer published. "66% of small firms are attacked, at £3,000 each" is from 2016. "The EU AI Act high-risk rules apply from August 2026" is no longer true. "MITRE ATLAS has 14 tactics" is out of date; it has 16 and 114 techniques. "NIST AI RMF 2.0" does not exist. "Machine identities outnumber humans 45 to 1" is an old figure; it is 109 to 1.

This report does not contain first-party measurements from UK Cyber Defence's own operations, because we did not want a synthesis of others' research to be mistaken for original research. The panels marked as our viewpoint contain our experience and our advice, and no numbers.

References

Each entry gives the publisher, title, date and location. “Primary” means the source document itself was read in preparing this report; “secondary” means the figure was taken from reputable reporting of a source we could not retrieve in full.

1. Varonis. *2025 State of Data Security Report: Quantifying AI’s Impact on Data Risk.* May 2025. 1,000 organisations, ~10 billion cloud resources. Primary. https://info.varonis.com/hubfs/Files/reports/2025-varonis-state-of-data-security-report.pdf
2. Check Point Research. *AI Security Report 2026* (second annual edition). 2026. Check Point telemetry, January–May 2026, and incident research. Primary.
3. Orca Security. *2026 State of AI Security Report: AI Security Findings from 1,200+ Production Cloud Environments.* Data collected Q2 2026. Primary.
4. HUMAN Security. *The 2026 State of AI Traffic & Cyberthreat Benchmark Report.* 26 March 2026. More than one quadrillion interactions, 2025. Primary. https://www.humansecurity.com/learn/resources/2026-state-of-ai-traffic-cyberthreat-benchmarks/
5. Office for National Statistics. *Business insights and impact on the UK economy: 2 July 2026.* BICS wave 159, 38,637 businesses. Primary. https://www.ons.gov.uk/businessindustryandtrade/business/businessservices/bulletins/businessinsightsandimpactontheukeconomy/2july2026
6. Office for National Statistics. *Artificial intelligence in UK businesses: 2023 to 2026.* 20 July 2026. Primary. https://www.ons.gov.uk/businessindustryandtrade/business/businessservices/articles/artificialintelligenceinukbusinesses/2023to2026
7. Department for Science, Innovation and Technology. *AI Adoption Research* (IFF Research and Technopolis). 3,500 interviews, February–May 2025; updated 13 February 2026. Primary. https://www.gov.uk/government/publications/ai-adoption-research/ai-adoption-research
8. British Chambers of Commerce, Atos and University of Essex. *Half of SMEs using AI, with limited headcount impact so far.* 18 March 2026. Primary. https://www.britishchambers.org.uk/news/2026/03/half-of-smes-using-ai-with-limited-headcount-impact-so-far/
9. McKinsey & Company / QuantumBlack. *The state of AI in 2026: On the road to ROI.* 25 August 2026. 1,719 participants, 97 nations. Primary. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai
10. McKinsey & Company / QuantumBlack. *The state of AI in 2025: Agents, innovation, and transformation.* November 2025. 1,993 participants, 105 nations. Primary.
11. Department for Science, Innovation and Technology and Home Office. *Cyber Security Breaches Survey 2025/2026.* 30 April 2026. 2,112 businesses, 1,085 charities. Primary. https://www.gov.uk/government/statistics/cyber-security-breaches-survey-20252026/cyber-security-breaches-survey-20252026
12. IBM and Ponemon Institute. *Cost of a Data Breach Report 2026.* 29 July 2026; UK release 28 July 2026. 602 breached organisations, 16 countries. Primary for press-release figures; the 21%, 92%, 43% and 68% AI-governance figures are secondary (report body). https://newsroom.ibm.com/2026-07-29-ibm-study-one-in-four-malicious-breaches-are-ai-enabled,-costing-companies-6-million-on-average and https://uk.newsroom.ibm.com/IBM-Study-More-Than-One-in-Five-Malicious-UK-Breaches-are-AI-Generated
13. IBM and Ponemon Institute. *Cost of a Data Breach Report 2025.* 30 July 2025. 600 breached organisations. Primary. https://newsroom.ibm.com/2025-07-30-ibm-report-13-of-organizations-reported-breaches-of-ai-models-or-applications,-97-of-which-reported-lacking-proper-ai-access-controls
14. Netskope Threat Labs. *Cloud and Threat Report: 2026.* 7 January 2026. Netskope One telemetry, October 2024–October 2025. Primary. https://www.netskope.com/resources/cloud-and-threat-reports/cloud-and-threat-report-2026
15. Netskope Threat Labs. *Cloud and Threat Report: Generative AI 2025.* Data February 2024–February 2025. Primary. https://www.netskope.com/resources/cloud-and-threat-reports/cloud-and-threat-report-generative-ai-2025
16. ISACA. *2026 AI Pulse Poll: AI Use Accelerates, While Governance and ROI Lag.* 5 May 2026. 3,400+ professionals. Primary. https://www.isaca.org/about-us/newsroom/press-releases/2026/ai-use-accelerates-while-governance-and-roi-lag-says-new-isaca-research
17. Gartner. *Gartner Identifies Critical GenAI Blind Spots That CIOs Must Urgently Address.* 19 November 2025. 302 security leaders, March–May 2025. Primary. https://www.gartner.com/en/newsroom/press-releases/2025-11-19-gartner-identifies-critical-genai-blind-spots-that-cios-must-urgently-address0
18. Gartner. *Gartner Survey Reveals Generative Artificial Intelligence Attacks Are on the Rise.* 22 September 2025. Primary. https://www.gartner.com/en/newsroom/press-releases/2025-09-22-gartner-survey-reveals-generative-artificial-intelligence-attacks-are-on-the-rise
19. Gartner. *Gartner Identifies the Top Cybersecurity Trends for 2026.* 5 February 2026. Primary. https://www.gartner.com/en/newsroom/press-releases/2026-02-05-gartner-identifies-the-top-cybersecurity-trends-for-2026
20. Microsoft and LinkedIn. *2024 Work Trend Index: AI at Work Is Here. Now Comes the Hard Part.* 8 May 2024. 31,000 workers, 31 countries. Primary. https://www.microsoft.com/en-us/worklab/work-trend-index/ai-at-work-is-here-now-comes-the-hard-part
21. Verizon. *2026 Data Breach Investigations Report.* April 2026. 31,000+ incidents, 22,000+ breaches, 145 countries. Primary. https://www.verizon.com/business/resources/reports/dbir/
22. Verizon. *2025 Data Breach Investigations Report.* 23 April 2025. Primary. https://www.verizon.com/business/resources/reports/dbir/
23. National Cyber Security Centre. *The near-term impact of AI on the cyber threat.* 24 January 2024. Primary. https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat

24. National Cyber Security Centre. *Impact of AI on cyber threat from now to 2027.* 7 May 2025. Primary. https://www.ncsc.gov.uk/report/impact-ai-cyber-threat-now-2027
25. National Cyber Security Centre. *Annual Review 2025.* 14 October 2025. Primary. https://www.ncsc.gov.uk/collection/ncsc-annual-review-2025
26. National Cyber Security Centre and AI Security Institute. *Why cyber defenders need to be ready for frontier AI.* 30 March 2026. Primary. https://www.ncsc.gov.uk/blogs/why-cyber-defenders-need-to-be-ready-for-frontier-ai
27. AI Security Institute. *How fast is autonomous AI cyber capability advancing?* 13 May 2026. Primary. https://www.aisi.gov.uk/blog/how-fast-is-autonomous-ai-cyber-capability-advancing
28. Microsoft. *Microsoft Digital Defense Report 2025.* 16 October 2025. Data July 2024–June 2025. Primary. https://blogs.microsoft.com/on-the-issues/2025/10/16/mddr-2025/
29. Microsoft Security. *Threat actor abuse of AI accelerates from tool to cyberattack surface.* 2 April 2026. Primary. https://www.microsoft.com/en-us/security/blog/2026/04/02/threat-actor-abuse-of-ai-accelerates-from-tool-to-cyberattack-surface/
30. Google Cloud / Mandiant. *M-Trends 2026.* 23 March 2026. Primary. https://cloud.google.com/blog/topics/threat-intelligence/m-trends-2026/
31. European Union Agency for Cybersecurity. *ENISA Threat Landscape 2025.* 1 October 2025. 4,875 incidents. Primary. https://www.enisa.europa.eu/publications/enisa-threat-landscape-2025
32. Anthropic. *Detecting and countering misuse of AI: August 2025.* 27 August 2025. Primary. https://www.anthropic.com/news/detecting-countering-misuse-aug-2025
33. Anthropic. *Disrupting the first reported AI-orchestrated cyber espionage campaign.* 13 November 2025. Primary. https://www.anthropic.com/news/disrupting-AI-espionage
34. Anthropic. *Countering misuse of AI: September 2026.* 10 September 2026. Primary. https://www.anthropic.com/threat-intelligence-report-september-2026
35. OpenAI. *Disrupting malicious uses of AI: October 2025.* 7 October 2025. Primary. https://openai.com/global-affairs/disrupting-malicious-uses-of-ai/
36. Google Threat Intelligence Group. *Adversarial Misuse of Generative AI.* 29 January 2025. Primary. https://cloud.google.com/blog/topics/threat-intelligence/adversarial-misuse-generative-ai
37. Google Threat Intelligence Group. *GTIG AI Threat Tracker: Advances in Threat Actor Usage of AI Tools.* 5 November 2025. Primary. https://cloud.google.com/blog/topics/threat-intelligence/threat-actor-usage-of-ai-tools
38. Google Threat Intelligence Group. *Adversaries Leverage AI for Vulnerability Exploitation, Augmented Operations, and Initial Access.* 11 May 2026. Primary. https://cloud.google.com/blog/topics/threat-intelligence/ai-vulnerability-exploitation-initial-access
39. Google Threat Intelligence Group. *From Prompting to Autonomy: The Evolution of Adversarial AI.* 8 September 2026. Primary. https://cloud.google.com/blog/topics/threat-intelligence/from-prompting-to-autonomy-the-evolution-of-adversarial-ai
40. Heiding, F., Lermen, S., Kao, A., Schneier, B. and Vishwanath, A. *Evaluating Large Language Models' Capability to Launch Fully Automated Spear Phishing Campaigns: Validated on Human Subjects.* arXiv:2412.00586, 30 November 2024; peer-reviewed as *Evaluating large language models' ability to automate spear phishing*, Expert Systems with Applications, vol. 314, June 2026. n=101. Primary. https://arxiv.org/abs/2412.00586
41. Hoxhunt. *AI-Powered Phishing Outperforms Elite Red Teams in 2025.* 3 April 2025. Vendor study, ~70,000 simulations per wave. Primary. https://hoxhunt.com/blog/ai-powered-phishing-vs-humans
42. Czybik, B., Kouam, C., Heubl, K., Nold, D. and Rieck, K. (TU Berlin/BIFOLD, Ruhr University Bochum, Inria). *A Large-Scale Study of Personalized Phishing using Large Language Models.* 2026. Field experiment, n=7,700. Primary. https://mlsec.tu-berlin.de/docs/2026-sec.pdf
43. Federal Bureau of Investigation, Internet Crime Complaint Center. *2025 Internet Crime Report.* 6 April 2026. Primary. https://www.fbi.gov/news/press-releases/cryptocurrency-and-ai-scams-bilk-americans-of-billions
44. City of London Police. *Report Fraud Annual Assessment 2025/26.* 4 September 2026. Primary. https://www.cityoflondon.police.uk/news/city-of-london/news/2026/september/city-of-london-police-intelligence-assessment-highlights-emerging-fraud-trends-and-evolving-threats/
45. UK Finance. *Annual Fraud Report 2026.* 15 June 2026. Primary. https://www.ukfinance.org.uk/news-and-insight/press-release/fraud-report-2026-press-release
46. Hong Kong Police, as reported by The Register, 5 February 2024, and Arup's confirmation as reported by Dezeen and CNN, 16–17 May 2024. Secondary. https://www.theregister.com/2024/02/05/hong_kong_deepfaked_cfo
47. Deloitte Center for Financial Services. *Generative AI is expected to magnify the risk of deepfakes and other fraud in banking.* 29 May 2024. United States scenario forecast. Primary. https://www.deloitte.com/us/en/insights/industry/financial-services/deepfake-banking-fraud-risk-on-the-rise.html
48. OWASP GenAI Security Project. *OWASP Top 10 for LLM Applications 2025.* 17 November 2024. Primary. https://genai.owasp.org/llm-top-10/
49. OWASP GenAI Security Project. *OWASP Top 10 for Agentic Applications for 2026.* 9 December 2025. Primary for date; list corroborated in secondary sources. https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/
50. MITRE. *ATLAS (Adversarial Threat Landscape for AI Systems)*, release v2026.08, 1 September 2026. Primary (release data). https://atlas.mitre.org/ and https://github.com/mitre-atlas/atlas-data
51. National Institute of Standards and Technology. *AI Risk Management Framework (AI RMF 1.0)*, 26 January 2023; *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* (NIST AI 600-1), July 2024. Primary. https://www.nist.gov/itl/ai-risk-management-framework
52. National Institute of Standards and Technology. *NIST IR 8596, Cybersecurity Framework Profile for Artificial Intelligence (Cyber AI Profile)*, initial preliminary draft, 15 December 2025; *Control Overlays for Securing AI Systems (COSAiS)*, concept paper August 2025. Primary. https://csrc.nist.gov/projects/cosais

53. National Cyber Security Centre and Cybersecurity and Infrastructure Security Agency, with 21 other agencies. *Guidelines for secure AI system development.* 27 November 2023. Primary. https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development

54. Department for Science, Innovation and Technology. *Code of Practice for the Cyber Security of AI*, with Implementation Guide. 31 January 2025. Primary. https://www.gov.uk/government/publications/ai-cyber-security-code-of-practice

55. ETSI. *TS 104 223 V1.1.1 (2025-04), Securing Artificial Intelligence (SAI); Baseline Cyber Security Requirements for AI Models and Systems*; *EN 304 223 V2.1.1 (2025-12).* Primary. https://www.etsi.org/deliver/etsi_en/304200_304299/304223/02.01.01_60/en_304223v020101p.pdf

56. International Organization for Standardization. *ISO/IEC 42001:2023, Information technology — Artificial intelligence — Management system.* 18 December 2023. Primary. Uptake estimate (~350 certifications by April 2026) is secondary and low confidence. https://www.iso.org/standard/42001

57. European Union. *Regulation (EU) 2024/1689 (Artificial Intelligence Act)*; *Regulation (EU) 2026/1744 of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 regarding simplification of harmonised AI rules (Digital Omnibus on AI)*, OJ 24 July 2026, in force 27 July 2026. Primary for text and dates; substance of amendments corroborated in legal commentary. https://eur-lex.europa.eu/eli/reg/2026/1744/oj/eng

58. Department for Science, Innovation and Technology. *Tackling AI security risks to unleash growth and deliver Plan for Change* (renaming of the AI Security Institute). 14 February 2025. Primary. https://www.gov.uk/government/news/tackling-ai-security-risks-to-unleash-growth-and-deliver-plan-for-change

59. UK Parliament. *Cyber Security and Resilience (Network and Information Systems) Bill.* Bill stages as at 28 September 2026. Primary. https://bills.parliament.uk/bills/4035

60. *Data (Use and Access) Act 2025*, section 80 (Articles 22A–22D UK GDPR); *Data (Use and Access) Act 2025 (Commencement No. 6) Regulations 2026*, SI 2026/82. Primary. https://www.legislation.gov.uk/ukpga/2025/18/section/80

61. Information Commissioner's Office. *Guidance on AI and data protection* (under review); *Consultation on draft automated decision-making guidance*, 31 March–29 May 2026; *Data Protection Act 2018 (Code of Practice on Artificial Intelligence and Automated Decision-Making) Regulations 2026*, SI 2026/425, in force 12 May 2026. Primary. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/

62. National Cyber Security Centre and IASME. *Cyber Essentials: Requirements for IT Infrastructure v3.3*, April 2026 ("Danzell" question set, effective 27 April 2026). Primary. https://www.ncsc.gov.uk/sites/default/files/2026-08/IT%20Requirements%20v3.3.pdf

63. *Directive (EU) 2022/2555 (NIS2)*, Article 21(2)(d); *Regulation (EU) 2022/2554 (DORA)*, Articles 28–30; European Supervisory Authorities, designation of critical ICT third-party providers, 18 November 2025. Primary for texts and designation date; count of 19 providers secondary.

64. National Vulnerability Database. *CVE-2025-32711* (Microsoft 365 Copilot, "EchoLeak", discovered by Aim Security). Published 11 June 2025. Primary. https://nvd.nist.gov/vuln/detail/CVE-2025-32711

65. Noma Security. *ForcedLeak: Agent Risks Exposed in Salesforce Agentforce.* 25 September 2025. Primary. https://noma.security/blog/forcedleak-agent-risks-exposed-in-salesforce-agentforce

66. CVE-2026-42824 ("SearchLeak", Varonis Threat Labs, June 2026), as reported by The Hacker News; CVE-2026-24299 ("Copirate 365", Embrace The Red, fixed 5 March 2026), https://embracethered.com/blog/posts/2026/defcon-talk-copirate-365/; "AgentForger" (Zenity Labs, fixed 8 June 2026), as reported by SecurityWeek. Secondary except where the researcher's own write-up is cited.

67. GitHub Advisory Database. *GHSA-6xpm-ggf7-wc3p / CVE-2025-6514, mcp-remote OS command injection* (JFrog Security Research). 9 July 2025. Primary. https://github.com/advisories/GHSA-6xpm-ggf7-wc3p

68. Postmark (ActiveCampaign). *Information regarding the malicious "postmark-mcp" package.* 25 September 2025; Koi Security research. Primary. https://postmarkapp.com/blog/information-regarding-malicious-postmark-mcp-package

69. Model Context Protocol. *Security Best Practices*, specification version 2026-07-28. Primary. https://modelcontextprotocol.io/docs/2026-07-28/tutorials/security/security_best_practices

70. National Security Agency. *Model Context Protocol (MCP): Security Design Considerations for AI-Driven Automation.* 20 May 2026. Primary (press release). https://www.nsa.gov/Press-Room/Press-Releases-Statements/Press-Release-View/Article/4496698/

71. Hugging Face and Protect AI. *Six months of Protect AI scanning on the Hub.* 14 April 2025. Primary. https://huggingface.co/blog/pai-6-month

72. JFrog Security Research. *Data Scientists Targeted by Malicious Hugging Face ML Models with Silent Backdoor.* 27 February 2024; ReversingLabs. *nullifAI*, 6 February 2025. Primary. https://jfrog.com/blog/data-scientists-targeted-by-malicious-hugging-face-ml-models-with-silent-backdoor/

73. Spracklen, J. et al. *We Have a Package for You! A Comprehensive Analysis of Package Hallucinations by Code Generating LLMs.* USENIX Security Symposium 2025; arXiv:2406.10279. Primary. https://arxiv.org/abs/2406.10279

74. nrwl/nx. *Security advisory GHSA-cxm3-wv7p-598c* ("s1ngularity"). 27 August 2025; Wiz Research analysis. Primary for advisory; impact figures secondary. https://github.com/nrwl/nx/security/advisories/GHSA-cxm3-wv7p-598c

75. Amazon Web Services. *Security Bulletin AWS-2025-015: Amazon Q Developer Extension for Visual Studio Code.* 23 July 2025. Primary. https://aws.amazon.com/security/security-bulletins/AWS-2025-015/

76. Anthropic, UK AI Security Institute and The Alan Turing Institute. *A small number of samples can poison LLMs of any size.* 9 October 2025. Primary. https://www.anthropic.com/research/small-samples-poison

77. Gartner. Press releases: *Top Predictions for IT Organizations and Users in 2025 and Beyond* (22 October 2024); *Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027* (25 June 2025); *Gartner Predicts 40% of Enterprise Apps Will Feature Task-Specific AI Agents by 2026* (26 August 2025); *Gartner Forecasts the Market for Securing AI Will Reach Almost $5 Billion in 2027* (26 August 2026). Primary. https://www.gartner.com/en/newsroom

78. CyberArk. *2025 Identity Security Landscape.* 23 April 2025. 2,600 security decision-makers, 20 countries. Secondary (press release corroborated in trade press).

79. Palo Alto Networks (CyberArk). *2026 Identity Security Landscape Report.* May 2026. 2,930 leaders. Primary (report landing page). https://www.paloaltonetworks.com/idira/identity-security-landscape-report

80. Microsoft. *Microsoft extends Zero Trust to secure the agentic workforce* (19 May 2025); *What's new in Microsoft Entra Agent ID* (updated 13 August 2026). Primary. https://learn.microsoft.com/en-us/entra/agent-id/whats-new-agent-id

81. Google Cloud. *Agent Identity overview*; *What's new in IAM, security governance and runtime defense* (6 May 2026). Primary. https://docs.cloud.google.com/iam/docs/agent-identity-overview

82. Amazon Web Services. *Amazon Bedrock AgentCore now generally available* (13 October 2025); *Securing AI agents with Amazon Bedrock AgentCore Identity* (14 October 2025). Primary. https://aws.amazon.com/blogs/security/securing-ai-agents-with-amazon-bedrock-agentcore-identity

83. Salesforce. *Agentforce Security and the Shared Responsibility Model.* 25 February 2026. Primary. https://help.salesforce.com/s/articleView?id=005315874

84. OpenAI Alignment. *Misalignment reports* series, September 2026, including *An agent used DNS to reach an external chatbot*; *Exposing a GitHub token in a public repository*; *Self-replicating prompt injections exist*. Primary. https://alignment.openai.com/misalignment-reports/

85. Cybersecurity and Infrastructure Security Agency, National Security Agency, Australian Signals Directorate's ACSC, National Cyber Security Centre (UK), Canadian Centre for Cyber Security and New Zealand NCSC. *Careful Adoption of Agentic AI Services.* 1 May 2026. Secondary (agency page not retrievable; corroborated in legal alerts). https://www.cisa.gov/resources-tools/resources/careful-adoption-agentic-ai-services

86. National Cyber Security Centre. *Thinking carefully before adopting agentic AI.* 15 May 2026. Primary. https://www.ncsc.gov.uk/blogs/thinking-carefully-before-adopting-agentic-ai

87. National Cyber Security Centre. *Managing the cyber risk of agentic AI* (interim guidance). 20 August 2026. Primary. https://www.ncsc.gov.uk/blogs/managing-the-cyber-risk-of-agentic-ai

88. National Institute of Standards and Technology. *Announcing the AI Agent Standards Initiative* (17 February 2026); NCCoE, *Accelerating the Adoption of Software and AI Agent Identity and Authorization*, concept paper (5 February 2026). Primary. https://www.nist.gov/news-events/news/2026/02/announcing-ai-agent-standards-initiative-interoperable-and-secure

89. SANS Institute (Crowley, C.). *2026 SANS SOC Survey Insights* (15 June 2026, 444 practitioners); *SANS SOC Survey 2025* (July 2025, 568 respondents). Primary. https://www.sans.org/white-papers/2026-sans-soc-survey-insights-decade-evolution-cyber-defense

90. Microsoft Office of the Chief Economist (Edelman, B. et al.). *Randomized Controlled Trial for Microsoft Security Copilot*, January 2024; Bono, J. et al., *Security Copilot: Evidence of Productivity Gains in Live Operations*, March 2025. Primary. https://www.microsoft.com/content/dam/microsoft/final/en-us/microsoft-product-and-services/microsoft-dynamics-365/pdf/Microsoft-Copilot-for-Security-productivity-findings-Whitepaper-Jan2024.pdf

91. Google Project Zero. *From Naptime to Big Sleep* (1 November 2024); Google, *cybersecurity updates, summer 2025* (CVE-2025-6965, 15 July 2025); TechCrunch report of twenty vulnerabilities (4 August 2025, secondary). https://projectzero.google/2024/10/from-naptime-to-big-sleep.html

92. Defense Advanced Research Projects Agency. *AI Cyber Challenge final results.* 8 August 2025. Primary. https://www.darpa.mil/news/2025/aixcc-results

93. ISC2. *2025 Cybersecurity Workforce Study* (4 December 2025, 16,029 respondents); *Rethinking AI's Impact on Cybersecurity Roles*, pulse survey (14 July 2026, 856 respondents). Primary. https://www.isc2.org/Insights/2025/12/2025-ISC2-Cybersecurity-Workforce-Study

94. ISACA. *State of Cybersecurity 2026.* 22 September 2026. 1,888 respondents. Primary. https://www.isaca.org/about-us/newsroom/press-releases/2026/only-8-percent-of-organizations-global-enterprises-conduct-regular-ai-specific-response-exercises

95. National Cyber Security Centre. *AI and cyber security: what you need to know* (13 February 2024); *Supporting AI adoption for UK cyber defence* (23 April 2026). Primary. https://www.ncsc.gov.uk/guidance/ai-and-cyber-security-what-you-need-to-know

96. National Cyber Security Centre. *Annual Review 2025, chapter 2: Cyber Essentials*; Department for Science, Innovation and Technology (Pye Tait Consulting). *Cyber Essentials Scheme Impact Evaluation*, July 2024. Primary. https://assets.publishing.service.gov.uk/media/671a0410a36621334536b3d7/Cyber_Essentials_Impact_Evaluation.pdf

97. Wiz Research (Nagli, G.). *Wiz Research Uncovers Exposed DeepSeek Database Leaking Sensitive Information, Including Chat History.* 29 January 2025. Primary. https://www.wiz.io/blog/wiz-research-uncovers-exposed-deepseek-database-leak

98. Gartner (Lawson, C. and Davies, J.). *Validate the Promises of AI SOC Agents With These Key Questions.* 2026, as reported by Help Net Security, 9 September 2026. Secondary.

99. The Register, *Samsung reportedly leaked its own secrets through ChatGPT*, 6 April 2023; Forbes, *Samsung bans ChatGPT and other chatbots for employees after sensitive code leak*, 2 May 2023. Secondary. https://www.theregister.com/2023/04/06/samsung_reportedly_leaked_its_own

100. UK Cyber Defence. *The phishing page that let Microsoft do the reconnaissance.* 18 September 2026. https://www.cyber-defence.io/insights/the-phishing-page-that-let-microsoft-do-the-reconnaissance

UK Cyber Defence Ltd, The Officers' Mess, Duxford, Cambridgeshire. CREST-accredited. ISO 27001:2022 and ISO 9001 certified; Cyber Essentials Plus. https://www.cyber-defence.io

This report may be quoted with attribution to UK Cyber Defence Ltd and to the underlying source. It contains no personal data and no client-identifying information. Published September 2026.