Safe Instinct

WHITEPAPER · 2026

The Knowledge Gap

Why 62% of breaches still run through employees, what security awareness training actually achieves, and how to build a human-risk programme that survives evidence.

62%

of breaches involve the human element

$5.29M

average cost of a breach that begins with a voice or text lure

6 mo

until trained phishing-detection skill returns to baseline

An evidence review for security leaders

safeinstinct.com

CONTENTS

What this paper covers

This is a research review, not a product brochure. Every figure in it is attributed to a named source with a date, and where the published evidence contradicts conventional practice — including practice common in the security awareness industry — the paper says so plainly.

NOTE ON METHOD AND SOURCING

Figures are drawn from primary reports wherever the primary document was directly accessible — the Verizon *2026 Data Breach Investigations Report*, the FBI IC3 *2025 Internet Crime Report*, the IBM/Ponemon *Cost of a Data Breach Report 2026*, peer-reviewed papers from IEEE Security & Privacy and USENIX, and the text of the regulations cited. Vendor telemetry is labelled as vendor telemetry. Where a widely-circulated statistic could not be traced to a checkable primary source, it has been left out rather than repeated. Several such figures are common in this market; their absence here is deliberate.

Superscript numbers refer to the reference list on page 24.

EXECUTIVE SUMMARY

The problem is not that employees don't care. It is that what we teach them does not survive contact with the attack.

Security awareness training is now near-universal, mandated by at least eight major regulatory regimes, and funded in almost every enterprise security budget. Over the same period, the share of breaches involving a human being has not fallen. In the 2026 Verizon dataset it rose — from 60% to 62%.[1] That gap between effort and outcome is the subject of this paper.

WHAT THE CURRENT EVIDENCE SHOWS

- **The human element is stable at roughly six in ten breaches**, and the attack surface around it has widened. Third-party involvement in breaches jumped from 30% to 48% in a single year, which means an organisation now inherits the knowledge gaps of its suppliers.[1]
- **The channel has shifted from the inbox to the phone line and the help desk.** In Mandiant's 2025 incident casework, voice phishing was the second most common initial infection vector at 11%, while email phishing fell to 6% — down from 22% in 2022.[7] Unit 42 found 36% of all incidents began with social engineering, and two-thirds of those targeted privileged accounts.[6]
- **The most rigorous studies of phishing training find little to no behavioural effect.** A randomised trial across 19,500 employees found no significant relationship between completing annual mandated training and clicking a phishing link; embedded "teachable moment" training reduced future clicking by about 2%.[14] An independent reproduction at a different firm, with 12,511 employees, found no significant improvement in either click rate or reporting rate.[15]
- **What training does achieve, it loses within six months.** Measured phishing-detection sensitivity roughly doubles immediately after training and is statistically indistinguishable from the pre-training baseline by month six — unless a reminder is delivered, which restores and sustains it.[17]
- **Cost is moving in the wrong direction.** The global average breach cost reached US$4.99M in 2026, and a breach that begins with a vishing or smishing lure averages $5.29M — the highest-cost entry point of any vector IBM measures.[3]
- **A new exposure class arrived faster than any training cycle.** Regular AI use on corporate devices tripled from 15% to 45% of employees in one year, 67% of it through non-corporate accounts — while 58% of workers report having received no security training on AI use at all.[1,11]

None of this argues for abandoning employee education. It argues that the dominant delivery model — an annual compliance module plus a punitive phishing simulation — is measuring completion rather than capability, and that the published evidence does not support the outcomes attributed to it. The organisations that reduce human risk treat it as a control domain with owners, telemetry, and a decay schedule, not as a training obligation with a due date.

Sections 1 through 4 establish the exposure. Section 5 examines what the peer-reviewed literature actually finds about awareness training, including the counter-evidence. Sections 6 through 9 set out a compliance baseline, seven design principles drawn from that evidence, a measurement set that replaces click rate, and a twelve-month sequence a security leader can put in front of a board.

"Anti-phishing training programs, in their current and commonly deployed forms, are unlikely to offer significant practical value in reducing phishing risks."

HO ET AL., IEEE SYMPOSIUM ON SECURITY AND PRIVACY, 2025 — RANDOMISED TRIAL, 19,500 EMPLOYEES

SECTION 01

The human element has not moved in five years

A stable statistic in a market that has spent heavily to change it.

The Verizon *Data Breach Investigations Report* is the closest thing the industry has to a longitudinal, multi-contributor breach census. Its 2026 edition analysed more than 31,000 security incidents and more than 22,000 confirmed breaches across 145 countries.[1] Its headline number for human involvement — errors, misuse, stolen credentials, and social engineering combined — was 62%, described in the report as a slight increase on the prior year's 60%.

This figure has hovered in the same band for half a decade. It has done so through a period in which security awareness training became a mandatory control under PCI DSS, NYDFS Part 500, DORA and NIS2; in which simulated phishing became standard practice; and in which the global market for awareness products grew substantially. Whatever the industry has been buying, it has not moved this number.

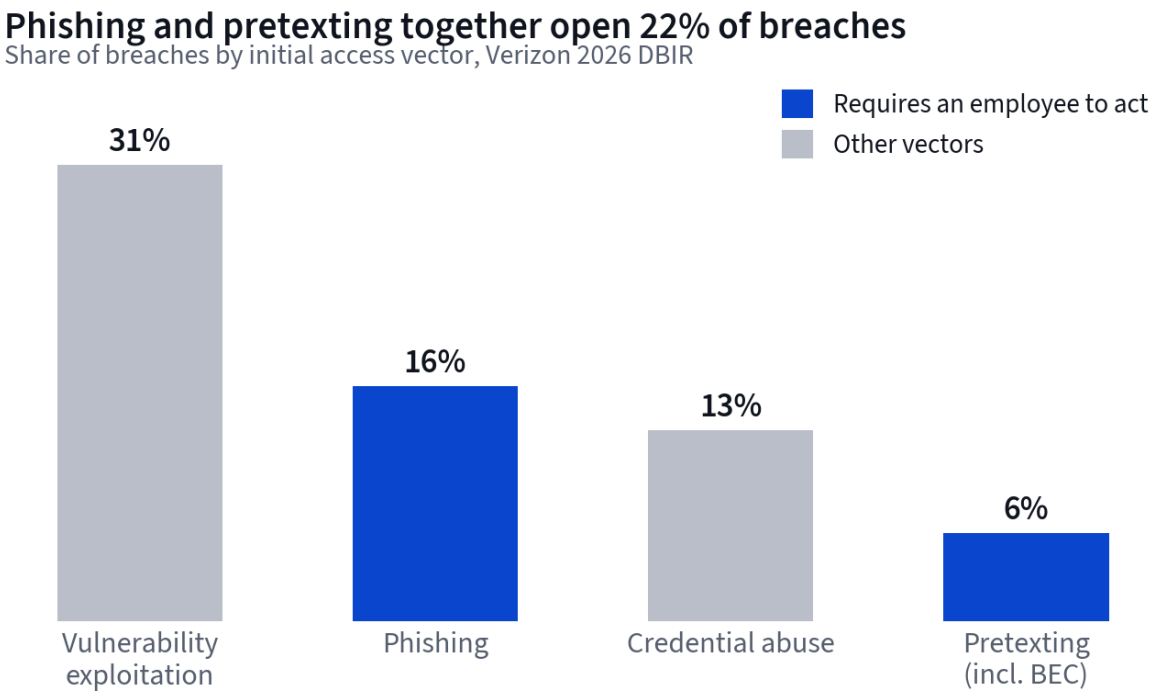


Figure 1. Phishing (16%) and pretexting (6%) — the two vectors that require an employee to take an action — together account for 22% of breach entry. Credential abuse is shown at its initial-access share of 13%; credentials appear somewhere in the chain of 39% of all breaches. Source: Verizon, 2026 DBIR.[1]

What changed underneath a stable number

The stability of "62%" conceals substantial movement in composition. Three shifts matter for anyone designing a human-risk programme.

Vulnerability exploitation overtook credential abuse as the leading initial access vector, rising from 20% to 31%.[1] This is often read as good news for the human-risk story — the machines are the problem now. It is not. Exploitation gets an attacker onto the network; it rarely gets them the data. In Mandiant's 2025 casework, the median time between an initial-access broker's entry and hand-off to a second threat actor fell from more than

eight hours in 2022 to 22 seconds in 2025.[7] Post-exploitation, the actor still has to move laterally, escalate, and persuade someone to approve something.

Third-party involvement in breaches rose from 30% to 48% — a 60% year-over-year increase.[1] Nearly half of breaches now involve a supplier, contractor, or integration partner. The practical consequence for a CISO is that the organisation's human-risk posture is now partly a function of training programmes it does not run and cannot audit directly. Ransomware, meanwhile, appeared in 48% of breaches, up from 44% — but ransomware is a monetisation model, not an entry method, and the entry is increasingly a person.[1]

62%	**48%**	**16%**	**6%**
of breaches involve the human element (2026, from 60%)	of breaches involve a third party (2026, from 30%)	of breaches begin with phishing — unchanged year on year	of breaches begin with pretexting, including BEC

The insurance market reached the same conclusion independently

Underwriters see a different slice of the same problem: not incidents, but paid losses. The cyber insurer Resilience reported that more than 85% of incurred cyber losses in the first half of 2026 stemmed from attacks that exploited human error, and that the share of incurred losses attributable to phishing, social engineering and transfer fraud rose from 17.7% in H1 2024 to 85.3% in H1 2026 — the largest single movement among loss drivers it tracks.[12]

Coalition's 2025 claims data points the same way from a different portfolio: business email compromise and funds-transfer fraud together accounted for 60% of all claims, with average BEC severity up 23% year on year to roughly $35,000, and 29% of BEC events escalating into a funds-transfer-fraud event averaging $106,000.[13]

Nearly half of breaches now involve a third party. An organisation's human-risk posture is increasingly a function of training programmes it does not run.

Why "the human element" is the wrong unit of analysis

The 62% figure is useful as a headline and almost useless as a management target, because it aggregates behaviours with nothing in common. A finance clerk approving a fraudulent payment after a deepfaked video call, a developer pasting a customer table into a public chatbot, a help-desk agent resetting MFA for an impersonator, and an administrator misconfiguring a storage bucket are four different failures with four different causes and four different fixes. Training all four with the same annual module is a category error.

Section 3 proposes a taxonomy that separates them. First, though, it is worth being precise about what employees are now facing — because the attack has changed considerably faster than the curriculum. It is also worth noting that insurers, who price risk they must pay for, have reached the same conclusion from two independent claims portfolios.

SECTION 02

How employee-targeted attacks changed, 2024–2026

Five shifts that most awareness curricula have not caught up with.

2.1 The channel moved from the inbox to the phone line

The most consequential change in this period is that the highest-value social engineering stopped arriving by email. Mandiant's *M-Trends 2026*, based on 2025 incident-response engagements, found voice phishing to be the second most common initial infection vector at 11% of investigations — behind exploits at 32%, but ahead of email phishing, which fell to 6% from 22% in 2022 and 14% in 2024.[7] CrowdStrike had recorded a 442% increase in vishing between the first and second halves of 2024.[8]

Two things drive this. Voice is far harder for technical controls to inspect than email — no gateway, no attachment sandbox, no URL rewriting on a phone call — and it is far better suited to the real-time, adaptive pressure that gets a person to override a process.

IBM's cost data reflects the difference in outcome: vishing and smishing produced the highest average cost of any entry point at $5.29M, against a global all-cause average of $4.99M.[3] Verizon's simulation telemetry shows the same asymmetry — mobile-centric lures achieved click success roughly 40% higher than equivalent email lures.[1] Almost no annual awareness curriculum trains for an inbound phone call. Almost every one trains for a suspicious email.

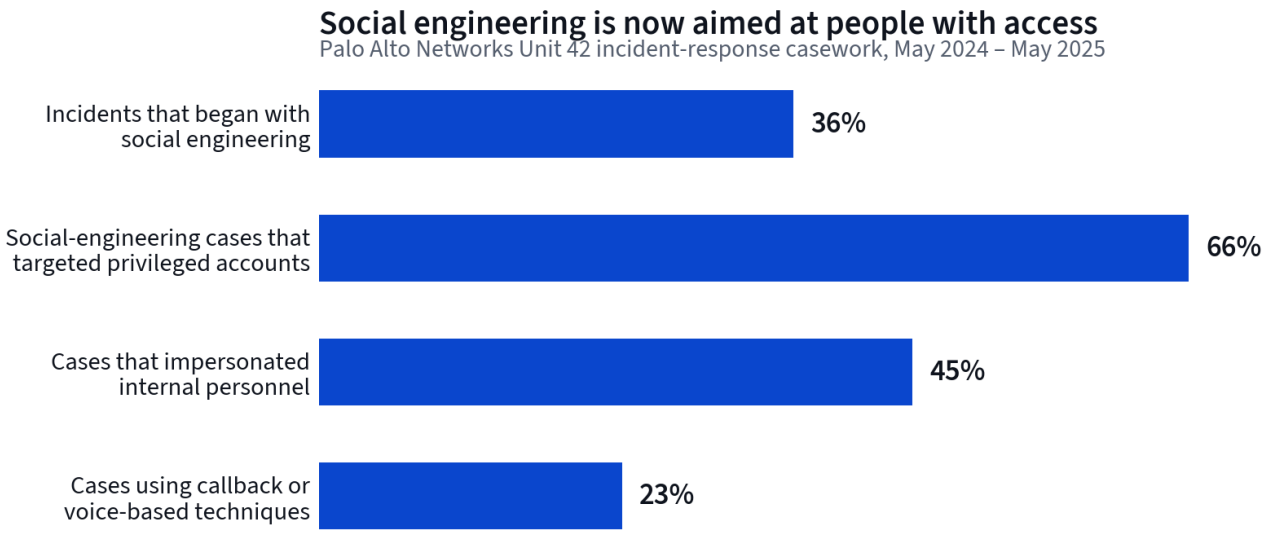


Figure 2. Unit 42 analysed a year of incident-response engagements and found social engineering to be the single most common starting point, at 36% of all incidents. Within those cases, impersonation of internal colleagues and targeting of privileged accounts dominate — a profile that points at help desks, IT staff and administrators rather than at the general employee population. Source: Unit 42, 2025 Global Incident Response Report: Social Engineering Edition.[6]

2.2 The help desk became an initial access vector

The clearest illustration is the 2025 UK retail campaign against Marks & Spencer, Co-op and Harrods. The UK National Cyber Security Centre's account of the tradecraft is unusually direct: attackers gathered employee

information from social media, executed SIM-swap fraud, then impersonated those employees to internal IT help desks and used intercepted authentication codes to complete password resets.[9] No malware was required. The control that failed was an identity verification procedure executed by a human being under time pressure. The financial consequence is public record: in its half-year results for the 26 weeks ended 27 September 2025, Marks & Spencer reported statutory profit before tax of £3.4M against £391.9M the prior year, including roughly £102M of one-off attack costs with a further £34M expected in the second half, partly offset by a £100M insurance payout.[10]

The same playbook recurred across US insurance and aviation through 2025, and in voice-phishing campaigns against Salesforce environments in which employees were persuaded by phone to authorise a malicious connected application into their organisation's tenant.[27]

WHAT THIS CHANGES FOR A SECURITY PROGRAMME

- The highest-risk population is not "all employees." It is the set of roles that can reset a credential, approve a payment, grant an OAuth consent, or change an entitlement.
- Identity verification at the help desk is a security control and should be engineered as one — with a scripted, non-overridable procedure, a callback to a known-good number, and an explicit, blame-free path to refuse a request.
- Teaching people to spot a suspicious email does not prepare them for a caller who knows their manager's name, their employee ID, and the project they shipped last week.

2.3 AI is an amplifier of social engineering, not a replacement for it

The 2026 DBIR's assessment is measured and worth quoting in substance: the median malicious actor observed using AI applied it across roughly 15 distinct documented attack techniques, and 44% of identified AI-assisted initial access was phishing-related — but under 2.5% of AI-assisted techniques were genuinely novel.[1] Attackers are not inventing new attacks with AI. They are producing existing attacks in greater volume, in more languages, with fewer of the errors that awareness training taught people to look for.

IBM found roughly one in four malicious breaches in its 2026 study were AI-enabled — a 56% increase year on year — at an average cost of about $6M, roughly $1M more than a comparable non-AI breach.[3] CrowdStrike recorded an 89% year-over-year increase in AI-enabled adversary operations and a 563% increase in incidents using fake-CAPTCHA lures.[28] Two incidents are unusually well documented: the 2024 Arup case, in which a finance employee in Hong Kong transferred approximately US$25M after a video conference in which the CFO and colleagues were deepfakes;[29] and Anthropic's November 2025 disclosure of a state-attributed campaign in which it assesses 80–90% of tactical operations were executed autonomously by an AI agent against roughly 30 target organisations.[30]

The implication for training content is narrow but important: the heuristics awareness programmes spent fifteen years teaching — poor grammar, odd formatting, generic salutations, an implausible sender — are now the weakest available signals. Guidance built on "spot the tell" is being actively obsoleted. Guidance built on "verify through a second channel you initiated" is not.

Under 2.5% of AI-assisted attack techniques observed in the 2026 DBIR were genuinely novel. AI is not producing new attacks; it is removing the flaws that awareness training taught people to look for.

VERIZON, 2026 DATA BREACH INVESTIGATIONS REPORT

2.4 Shadow AI arrived faster than any training cycle

Between the 2025 and 2026 DBIR editions, the proportion of employees regularly using generative AI on corporate devices rose from 15% to 45%. Of those users, 67% were accessing the tools through non-corporate accounts, placing the activity outside enterprise logging, retention and DLP. Shadow AI policy violations became the third most common non-malicious insider action, rising roughly fourfold.[1]

The independent survey evidence explains why. In the National Cybersecurity Alliance and CybSafe *Oh Behave!* study of more than 6,500 people across seven countries, 65% reported using AI tools, 58% had received no security or privacy training on AI use, and 43% admitted sharing sensitive workplace information — internal documents, financial data, client information — with an AI tool without employer approval.[11] Netskope recorded generative-AI data-policy violations more than doubling across 2025, averaging 223 per organisation per month.[31]

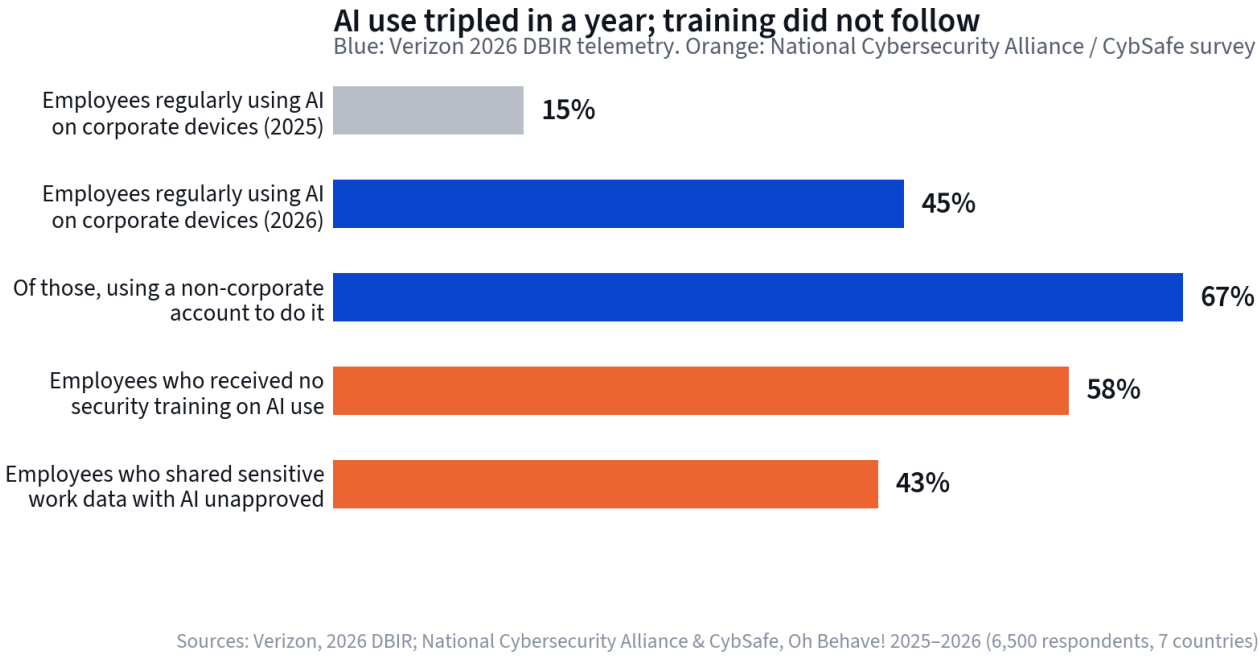


Figure 3. Adoption outran governance by roughly a full year — the clearest current example of a knowledge gap created faster than an annual training cycle can close it. Sources: Verizon 2026 DBIR;[1] National Cybersecurity Alliance & CybSafe.[11]

IBM quantifies the consequence: shadow AI figured in 43% of security incidents in its 2026 study, causing data loss in roughly half of those cases, and 92% of organisations lacked adequate access controls around their AI systems.[3] This is a useful test case for any human-risk programme, because it is a gap with a known creation date. An organisation that took twelve months to publish AI guidance was, by construction, twelve months late. Programmes designed on an annual cadence cannot respond to a capability that reaches half the workforce in a year.

2.5 Third parties inherit — and export — knowledge gaps

With third-party involvement in 48% of breaches,[1] the boundary of a human-risk programme no longer matches the boundary of the payroll. The 2025 Salesloft Drift incident is instructive: attackers obtained OAuth and refresh tokens for a widely-deployed integration and used them to exfiltrate data from customers' connected SaaS environments — exposing hundreds of organisations through a trust relationship none of their own employees had done anything wrong to create.[32]

The corresponding control is contractual, not educational: require evidence of an awareness and phishing-resistance programme from suppliers with production or identity access, and treat it as a reviewable artefact rather than a questionnaire checkbox. DORA takes this position explicitly.[34]

SECTION 03

A taxonomy of employee knowledge gaps

Seven distinct failure modes that a single annual module treats as one.

"Human error" is not a diagnosis. It is the point at which diagnosis stopped. The taxonomy below separates the employee-mediated failures that appear in the incident data into seven classes, each with a different root cause and therefore a different remedy. The value of the split is operational: it tells a security leader which gaps are training problems, which are process problems, and which are design problems that no amount of training will fix.

01 Recognition gap — the employee cannot identify the attack

The classic phishing-awareness target. The employee does not recognise a lure as hostile. This is the gap awareness training was built for, and it is the one that has degraded most as AI has removed the surface tells. It remains real — particularly for novel channels such as quishing, callback lures and fake-CAPTCHA "ClickFix" pages, where Microsoft found ClickFix to account for 47% of initial access incidents detected by its managed-response service.[33]

OBSERVABLE INDICATOR Click rate on lure families the workforce has not seen before; large spread in click rate across lure types.

02 Verification gap — the employee cannot safely confirm a request

The employee suspects something, but has no fast, sanctioned way to check. This is the gap that help-desk impersonation, deepfaked executive calls and payment fraud exploit. It is rarely a knowledge problem: most people know they should verify. They do not know *how* to verify in the next ninety seconds without appearing obstructive to someone senior.

OBSERVABLE INDICATOR Absence of a documented out-of-band verification procedure; median time to verify an unusual request; number of payment approvals with no second-channel confirmation.

03 Authority gap — the employee cannot refuse

The employee correctly identifies a request as suspicious and complies anyway, because refusing carries a social or career cost. Every deepfaked-CFO fraud depends on this gap. Training does not close it; explicit, senior, published permission to refuse does — as does removing the individual from the decision through dual authorisation.

OBSERVABLE INDICATOR Reported near-misses involving executive impersonation; whether a written "no one senior will ever penalise you for verifying" policy exists and is known.

04 Consequence gap — the employee does not understand what their access is worth

Common in help-desk, HR and finance roles, and the reason Unit 42 found 66% of social-engineering cases targeting privileged accounts.[6] The employee views a password reset or a vendor bank-detail change as an administrative task rather than as a security decision, because nothing in their training connected their routine action to a breach outcome.

OBSERVABLE INDICATOR Whether role-specific risk briefings exist for help desk, finance and HR; entitlement review findings showing standing privilege without corresponding training.

05 Currency gap — the guidance is older than the threat

The employee is doing exactly what they were taught, and what they were taught is obsolete. Shadow AI is the current worked example — 58% of workers report no training on AI use while 65% use it.[11] Two years ago the same gap existed for MFA push fatigue and QR-code lures.

> **OBSERVABLE INDICATOR** Elapsed time between a new attack technique appearing in threat intelligence and corresponding guidance reaching staff. If this exceeds one quarter, the gap is structural.

06 Reporting gap — the employee detects the attack and tells no one

Arguably the most expensive gap, because it converts a contained event into a dwell-time problem. The ETH Zurich longitudinal study found reporting rates varying by a factor of three across employee populations within a single company — 22% among frequent computer users against 7.6% among specialised users.[16] Punitive simulation programmes actively widen this gap by attaching a consequence to admitting a mistake.[22]

> **OBSERVABLE INDICATOR** Report rate as a proportion of delivered lures; ratio of reports to clicks; median time from delivery to first report.

07 Friction gap — the secure path is harder than the insecure one

Not a knowledge gap at all, but it presents as one. The employee uses a personal AI account because the sanctioned tool is slow; forwards a document to personal email because the file share is blocked; reuses a password because rotation policy made it unmanageable. NIST's security-fatigue research documents the resulting posture precisely: resignation, loss of control, and decision avoidance.[19] Training an employee out of a rational response to bad process is not possible.

> **OBSERVABLE INDICATOR** Volume of sanctioned-tool workarounds; proportion of AI use through non-corporate accounts (67% in the 2026 DBIR);[1] helpdesk tickets requesting exceptions.

Mapping gaps to the control that actually closes them

The reason to draw these distinctions is that only two of the seven are primarily training problems.

GAP	DOMINANT CONTROL	WHY TRAINING ALONE IS INSUFFICIENT
01 Recognition	Training + technical filtering	Genuine training target, but AI has removed the surface cues most curricula teach; content must be refreshed continuously, not annually.
02 Verification	Process design	People know they should verify. They need a defined, fast, sanctioned method and the time to use it.
03 Authority	Governance + dual control	A power asymmetry cannot be trained away. Remove the single point of human decision.
04 Consequence	Role-based briefing	Generic content cannot connect a specific action to a specific loss; only role-targeted material can.
05 Currency	Programme cadence	An annual cycle structurally guarantees a lag measured in quarters against a threat that moves in weeks.
06 Reporting	Culture + tooling	Reporting is suppressed by blame, not by ignorance. One-click reporting plus a no-fault stance moves it.
07 Friction	Engineering	Insecure workarounds are rational responses to unusable controls. Fix the control.

Gaps 01 and 04 respond to training. Gaps 02, 03, 05, 06 and 07 respond to process, governance, cadence, culture and engineering respectively. A programme that treats all seven as a content problem will under-deliver against five of them regardless of the quality of the content.

SECTION 04

Quantifying the exposure

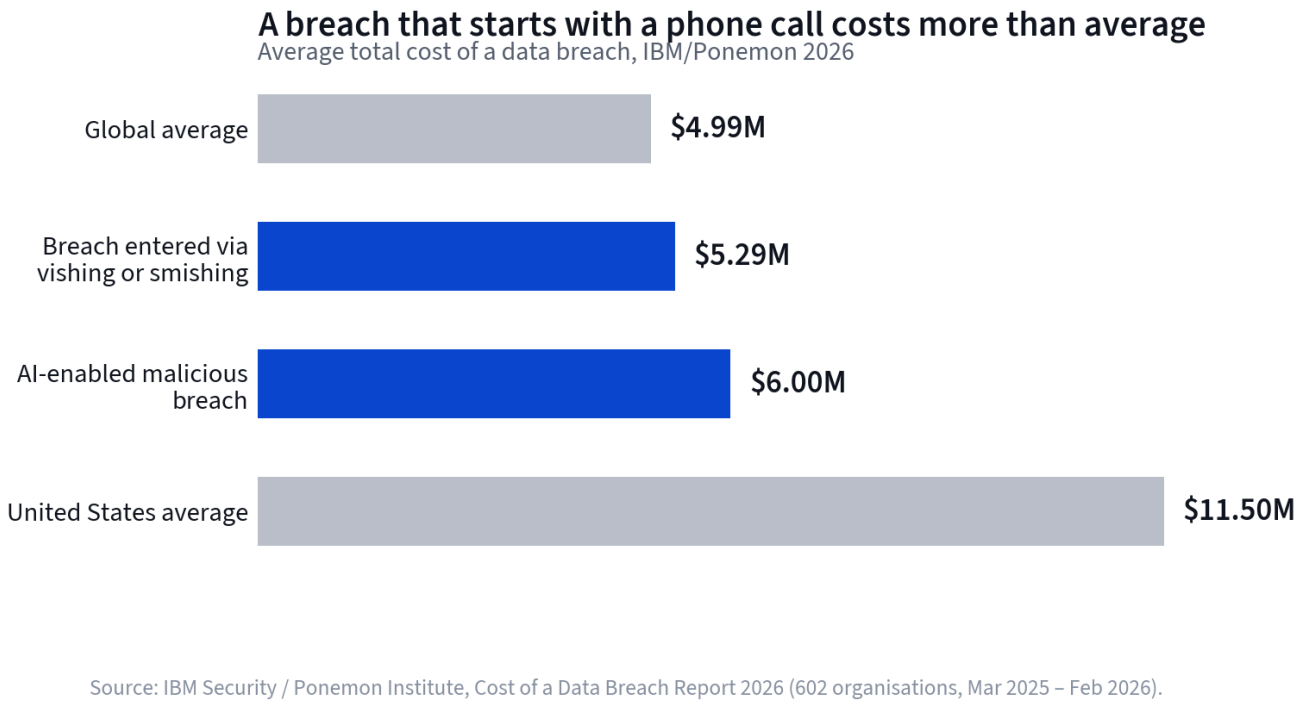
Putting a defensible number on human risk without inventing one.

Boards fund what they can size. The difficulty with human risk has always been that the credible loss data is aggregate while the organisation-specific data is thin. What follows is the published baseline, a method for turning it into an internal estimate that survives challenge, and an explicit account of what that method cannot tell you.

4.1 The published loss baseline

IBM and the Ponemon Institute's 2026 study of 602 organisations, covering incidents between March 2025 and February 2026, put the global average total cost of a data breach at US$4.99M — a 12% year-on-year increase and a record high — with the United States average at $11.5M.[3] Mean time to identify and contain rose to 247 days, reversing five consecutive years of improvement.

Phishing led all entry points for the fourth consecutive year. Vishing and smishing produced the highest average cost of any single vector at $5.29M.[3]



Source: IBM Security / Ponemon Institute, Cost of a Data Breach Report 2026 (602 organisations, Mar 2025 – Feb 2026).

Figure 4. Average total breach cost under different conditions. The two blue bars are human-mediated entry conditions; both exceed the global all-cause average. Source: IBM Security / Ponemon Institute, Cost of a Data Breach Report 2026.[3]

The FBI's Internet Crime Complaint Center recorded 1,008,597 complaints in 2025 with total reported losses of $20.877 billion. Business email compromise alone accounted for 24,768 complaints and $3.046 billion, with a further 191,561 complaints categorised as phishing or spoofing.[4] IC3 figures represent only what was reported to the FBI and are therefore a floor, not an estimate of true loss.

4.2 Insider risk: the majority of loss is not malicious

The Ponemon Institute and DTEX Systems 2026 insider-risk benchmark, covering 354 organisations and 7,490 incidents, puts the average annual cost of insider risk at US$19.5M globally and $24M in North America. Critically for programme design, 53% of incidents were attributed to negligent or mistaken insiders, 27% to malicious or criminal insiders, and 20% to credential theft.[5]

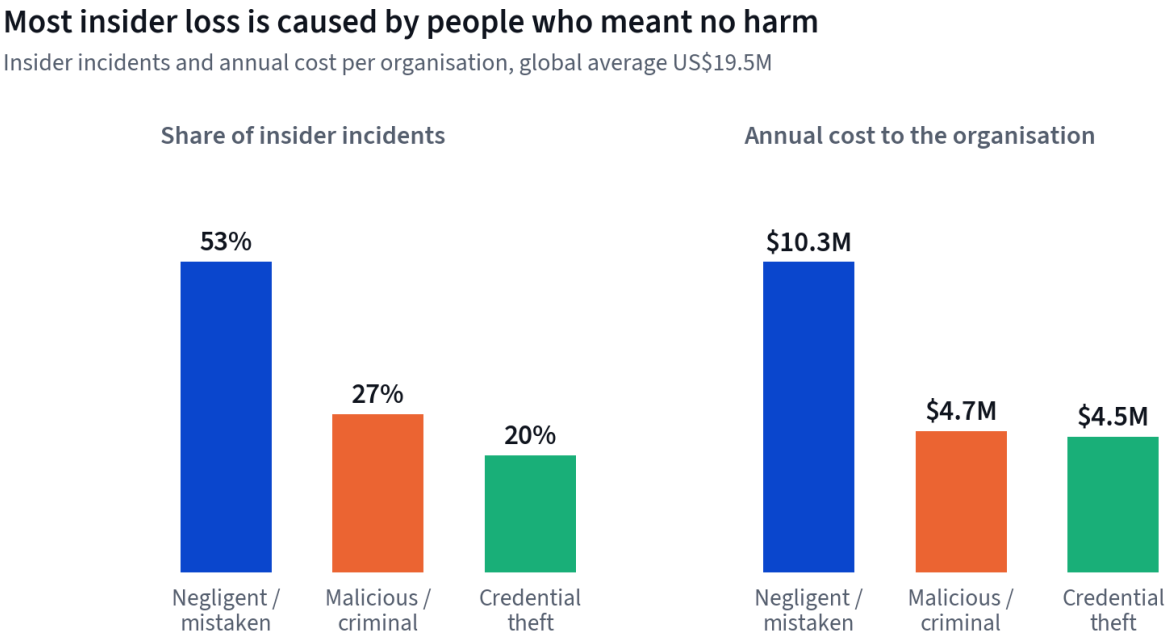


Figure 5. Negligence accounts for a majority of insider incidents and the largest share of annual cost. Average containment time improved to 67 days from 81; organisations containing incidents within 30 days incurred $14.2M annually against $21.9M for those taking more than 90 days. Source: Ponemon Institute / DTEX Systems, 2026 Cost of Insider Risks: Global Report.[5]

The containment differential is the most actionable number in that report: a $7.7M annual gap between fast and slow containment is largely a reporting-speed problem, addressable through the mechanisms in gap 06.

4.3 Building an internal estimate

The following structure produces a defensible internal figure. It is deliberately simple, because elaborate models invite argument about parameters rather than agreement about direction.

#	INPUT	WHERE IT COMES FROM
1	**Exposed population**	Headcount with access to systems or funds in scope, segmented by role. Not total headcount — the roles that can move money, reset credentials, grant consent, or export data.
2	**Encounter rate**	Observed hostile contacts per person per year from your own mail gateway, telephony and reporting data. Use internal telemetry; industry averages will be wrong for your sector.
3	**Failure rate**	Measured susceptibility from simulation and, more importantly, from real incidents. Segment by role: a single organisation-wide figure conceals the risk concentration.
4	**Containment factor**	Proportion of failures caught before material impact — a function of reporting rate and detection coverage. This is where a programme most visibly changes the number.
5	**Loss magnitude**	Scenario-specific. Use your own historical incidents where available; fall back to the published averages in §4.1 with the vector-specific figures, not the headline.

Annual expected loss ≈ (1 × 2 × 3) × (1 − 4) × 5. The output is a range, not a point. Present it as one.

WHAT THIS MODEL CANNOT DO — STATE THIS TO THE BOARD

It cannot establish that a given training investment caused a given reduction in loss. No published, independently audited study does this for security awareness training, and Section 5 explains why claims to the contrary should be read sceptically. The model sizes exposure and shows sensitivity to the containment factor; it is not a return-on-investment calculation, and presenting it as one invites a challenge the evidence cannot answer.

The honest framing is: *here is the size of the exposure, here is the single variable we can most reliably move (containment through reporting), and here is what we will measure to know whether we moved it.*

4.4 The concentration insight

Two independent datasets suggest human risk is concentrated rather than distributed. Verizon's 2025 analysis found that 8% of employees consistently account for the large majority of insecure data access and sharing incidents.[2] The ETH Zurich longitudinal study found that among employees who clicked at least one simulated phish, 30.6% went on to click two or more, and 23.9% took a dangerous action more than once.[16]

A programme that spends uniformly across the workforce spends most of its budget on people who were never going to fail, and under-serves the roles that carry the risk. Segmenting by measured susceptibility and by entitlement is the single highest-leverage change available to most programmes.

SECTION 05

What the evidence says about awareness training

The peer-reviewed literature is less flattering than the industry literature. Both are presented here.

Security awareness training is a mature market with a well-developed set of claims about its effectiveness, most of which originate from the organisations selling it. Over the last four years a body of independent, peer-reviewed field research has accumulated that tests those claims directly. A security leader deciding where to spend should know what it found.

5.1 The randomised trials find little to no behavioural effect

The most rigorous study to date was conducted at UC San Diego Health and published at the 2025 IEEE Symposium on Security and Privacy. Researchers embedded a randomised controlled trial inside the organisation's live phishing-training programme: approximately 19,500 employees, eight months, ten distinct simulated phishing campaigns.[14] Its findings:

- No significant relationship between having recently completed the annual mandated cybersecurity training and the likelihood of clicking a phishing link.
- Embedded "teachable moment" training, delivered immediately after a simulated click, reduced future click risk by approximately 2%.
- 75% of users who received embedded training engaged with the material for one minute or less; roughly one-third closed it without engaging at all.
- Susceptibility accumulated rather than declined: around 10% of users clicked in month one, and by month eight more than 50% had clicked at least once.
- Lure content dominated outcomes. An Outlook password-reset lure drew a 1.82% click rate; a vacation policy update drew 30.8%.

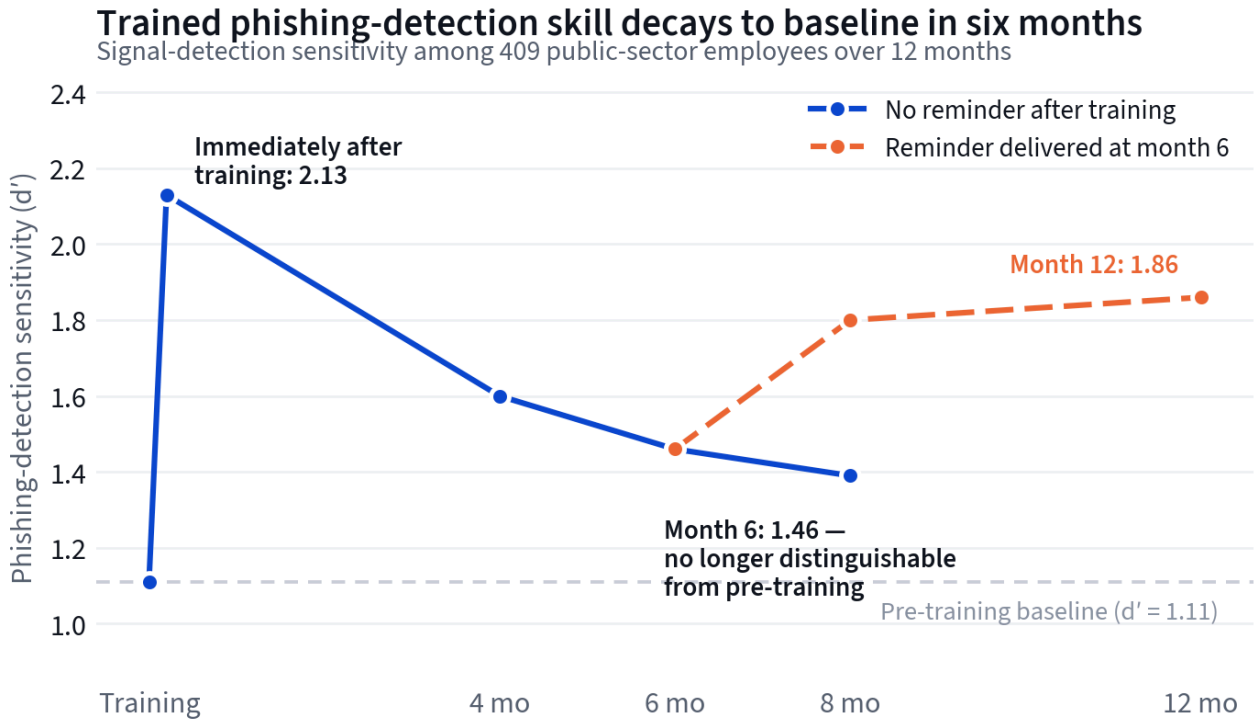
The authors' conclusion, stated in the paper, is that anti-phishing training in its current and commonly deployed forms is unlikely to offer significant practical value in reducing phishing risk.

This result has been independently reproduced. Researchers at Purdue ran a randomised trial across 12,511 employees at a US fintech firm, comparing lecture-only training, lecture-plus-interactive training, and no training, with lure difficulty controlled using the NIST Phish Scale. They found no statistically significant improvement in click rate ($p = 0.450$) or reporting rate ($p = 0.417$), with negligible effect sizes. Lure difficulty was the dominant predictor: click rates rose from 7.0% on easy lures to 15.0% on hard ones.[15]

Both align with the earlier ETH Zurich study — 14,733 employees, fifteen months, 117,864 simulated emails — which concluded that embedded training as commonly deployed does not make employees more resilient and can produce unintended side effects.[16]

5.2 What training does achieve decays within six months

The countervailing evidence is that training does produce a measurable short-term effect — it simply does not persist. A field study at a German state authority, published at USENIX SOUPS, tracked 409 employees for a full year using signal-detection sensitivity (d′) as the measure of phishing-detection skill.[17]



Source: Reinheimer et al., “An investigation of phishing awareness and education over time,” USENIX SOUPS 2020.

Figure 6. Detection sensitivity nearly doubles immediately after training, then decays; by month six it is no longer statistically distinguishable from the pre-training baseline. A reminder delivered at month six restores performance and sustains it through month twelve. The researchers recommend reminders at roughly four-month intervals. Source: Reinheimer et al., USENIX SOUPS 2020.[17]

This is the most useful single finding in the literature for programme design, because it converts a vague intuition ("training should be more frequent") into a specific parameter. An annual cadence leaves roughly half the year with no measurable residual effect. The decay curve, not the compliance calendar, should set the schedule.

5.3 Habituation and security fatigue are measurable, not metaphorical

Two independent research traditions explain why repetition alone does not solve decay. Functional MRI work presented at ACM CHI found a sharp drop in visual-processing activity after only the second exposure to an identical security warning, worsening with further repetition — and that polymorphic warnings, which vary their appearance, substantially resist this habituation.[18] Repeating the same message in the same format is neurologically self-defeating.

NIST's qualitative study of security fatigue found that more than half of participants raised fatigue spontaneously, without prompting, and documented the resulting posture: resignation, a sense of lost control, fatalism, risk minimisation and decision avoidance.[19] A workforce in this state is not under-informed. It is over-asked.

5.4 The knowledge–behaviour gap

Underlying all of this is a finding that runs through the usable-security literature: knowledge is a weak predictor of behaviour. A systematic review of 142 cybersecurity training studies found that while 62 of the empirical studies reported positive effects, the median sample was 96 participants, most tested training only once before measuring, most measured knowledge or intentions rather than behaviour, and only 26 of the 142 grounded their

design in an established behavioural theory.[20] The evidence base is not merely mixed; it is methodologically thin in exactly the place that matters.

Research on reporting behaviour points the same way. A 2025 cross-national study across Germany, the UK and the US found that the quality of security-related *communication* was a stronger predictor of phishing-reporting behaviour than knowledge was.[21]

Training produces a measurable effect that disappears in six months, on a threat that changes in six weeks, delivered on a twelve-month cycle. The arithmetic, not the content, is the problem.

5.5 The counter-evidence, stated fairly

It would be dishonest to present only the null results. Three lines of evidence point the other way.

Sustained, high-frequency programmes beat one-off training. A twelve-month study across 20 organisations and 1,300+ employees, using 13,000+ simulated emails, reported that sustained simulation with targeted training halved successful compromise rates within six months.[23] Note the design difference — continuous, not annual — which is consistent with the decay findings, not contrary to them.

Vendor benchmark data shows large susceptibility reductions. KnowBe4's 2026 benchmarking report, drawn from 42 million simulated exercises across 14.8 million employees at 64,000 organisations, reports a global baseline phish-prone percentage of 33.2%, falling to 20.1% after 90 days and 4.2% after twelve months.[24] This is by far the largest dataset available on the question, and it carries three caveats: it is vendor-produced and vendor-analysed; there is no untrained control group, so the effect of training cannot be separated from the effect of repeated exposure to simulations; and organisations still enrolled after a year are self-selected for commitment. The measured construct — susceptibility to that vendor's simulations — is also not susceptibility to real attacks.

Reporting, as distinct from detection, responds well. The ETH Zurich researchers found crowdsourced employee reporting to be an effective and sustainable mechanism for detecting live phishing campaigns, even as they found embedded training ineffective at reducing clicks.[16] This is the strongest positive finding in the independent literature, and it points at a different programme objective than the one most organisations pursue.

5.6 What this does and does not mean

None of this means awareness training is worthless or should be cut. It is a regulatory requirement in most sectors (Section 6), a prerequisite for several classes of technical control, and the counter-evidence above shows sustained programmes outperform episodic ones. It does mean four specific things:

1. **Budget annual compliance training as compliance, not as risk reduction.** The randomised evidence does not support a risk-reduction claim for it, and a claim the evidence cannot support is a liability in front of a board or a regulator.
2. **Click rate is a poor primary metric.** It is dominated by lure difficulty rather than workforce capability,[15] so a programme can improve it by choosing easier lures.
3. **Punitive simulation programmes are counter-productive.** Punishment discourages the reporting of real attacks — the behaviour with the strongest supporting evidence.[22] A USENIX Security study of a 16,000-employee campaign measured the cost directly: employees who clicked reported higher stress (5.40 of 10) and lower self-efficacy (3.56 of 5) than those who reported the email (3.39 and 4.1), while 83.6% still rated the campaign worthwhile — so buy-in is not what is being traded away.[25]
4. **Cadence and targeting matter more than content quality.** Given decay at six months and concentration in 8% of employees,[2,17] the same budget spent more often on fewer people beats it spent annually on everyone.

SECTION 06

The compliance floor — and why it is only a floor

Where employee training is legally required, and what regulators have actually enforced.

Security awareness training is now an explicit obligation across most major regulatory regimes a large organisation operates under. The table below gives the specific provision, what it requires, and its effective date, so that a programme can be mapped against it directly.

REGIME	PROVISION	WHAT IT REQUIRES	EFFECTIVE
EU NIS2 Directive 2022/2555	Art. 20(2); Art. 21(2)(g)	Members of management bodies must undergo training; entities are encouraged to offer similar training to staff regularly. Basic cyber hygiene and cybersecurity training is one of ten mandatory risk-management measures. Management bodies can be held personally liable for infringements.	Transposition due 17 Oct 2024
EU DORA Reg. 2022/2554	Art. 13(6)	ICT security awareness programmes and digital operational resilience training must be compulsory modules in staff training schemes, applying to all employees *and senior management*, and may be extended to ICT third-party providers.	Applies since 17 Jan 2025
EU AI Act Reg. 2024/1689	Art. 4	Providers and deployers must ensure a sufficient level of AI literacy among staff and others operating AI systems on their behalf, proportionate to technical knowledge, context and affected persons.	Applies since 2 Feb 2025
PCI DSS v4.0.1	12.6.3; 12.6.3.1; 12.6.3.2; 5.4.1	Awareness training on hire and at least every 12 months, with annual acknowledgement; content must explicitly include phishing, related attacks and social engineering, and acceptable use of end-user technologies. Automated mechanisms must protect personnel against phishing.	Future-dated requirements mandatory since 31 Mar 2025
NYDFS Part 500 23 NYCRR	§ 500.14(a)(3)	Periodic — at minimum annual — cybersecurity awareness training for all personnel, explicitly including social engineering, updated to reflect risks identified in the entity's risk assessment.	Training provision effective 29 Apr 2024
HIPAA Security Rule	45 CFR 164.308(a)(5)	A security awareness and training programme for all workforce members including management, with implementation specifications covering security reminders, malicious-software protection, log-in monitoring and password management. A proposed rule published 6 January 2025 would make currently addressable specifications mandatory; it had not been finalised as of July 2026.	In force; NPRM pending
ISO/IEC 27001:2022	Annex A 6.3; Cl. 7.2, 7.3	An ongoing awareness, education and training programme covering policy familiarity, personal accountability, baseline practices and event reporting. Clauses 7.2 and 7.3 require demonstrated competence and documented awareness for certification.	2022 edition
NIST CSF 2.0	PR.AT-01; PR.AT-02; GOVERN	General-workforce and specialised-role awareness and training so personnel can perform tasks with cybersecurity risk in mind. SP 800-50r1 (2024) replaces the awareness/training/education tiering with a unified learning continuum and adds guidance on measuring programme impact.	Voluntary framework, Feb 2024
GDPR	Art. 39(1)(b); Art. 32	The DPO must monitor compliance including awareness-raising and staff training. Article 32's "appropriate organisational measures" is the provision under which regulators have treated inadequate training as a security failure.	In force since 2018
SEC	Reg. S-K Item 106; Item 1.05 8-K	Annual disclosure of processes for assessing, identifying and managing material cybersecurity risks, board oversight, and management's role and expertise. Material incidents disclosed within four business days.	Effective Dec 2023

Regulatory status was verified as of July 2026. The HIPAA Security Rule NPRM and a pending petition to amend the SEC incident-disclosure requirement were both unresolved at that date; confirm current status before relying on either.

Regulators have enforced on training specifically

Two cases show that inadequate training is treated as a substantive failure rather than a paperwork one.

In October 2022 the UK Information Commissioner's Office fined **Interserve Group Ltd £4.4 million** after a phishing email led to a malware compromise exposing the personal data of up to 113,000 current and former employees. The ICO's published findings cited outdated software, inadequate monitoring, and a failure to properly train staff. Information Commissioner John Edwards's framing is worth noting: "The biggest cyber risk businesses face is not from hackers outside of their company, but from complacency within their company."[36]

In August 2025 the US Department of Health and Human Services Office for Civil Rights settled with **BST & Co. CPAs for $175,000** following a ransomware attack that began with phishing. Beyond the monetary settlement, the corrective action plan required security-awareness training tailored to the organisation and to workforce job duties — a regulator prescribing training design, not merely training existence.[26]

Cyber insurance treats it as an underwriting control

Awareness training and phishing simulation are now standard screening items on cyber insurance applications, alongside MFA, EDR and offline backups. Insurers ask whether a formal programme exists, at what frequency, and whether simulation testing is performed. Given the claims composition described in Section 1 — BEC and funds-transfer fraud at 60% of Coalition's claims,[13] human-error-driven attacks at over 85% of Resilience's incurred losses[12] — this is unlikely to relax.

WHY THE FLOOR IS ONLY A FLOOR

- Every regime above specifies *frequency* (annual) and *coverage* (all personnel). None specifies *efficacy*. A programme can be fully compliant and, on the evidence in Section 5, produce no measurable behavioural change.
- The minimum cadence these rules establish — annual — is the cadence the decay research identifies as insufficient.[17] Compliance and effectiveness point in different directions here, and the gap has to be closed deliberately.
- Where regulators have gone beyond frequency, they have gone to design and competency — as OCR did in the BST corrective action plan. That is the direction of travel worth planning for.

SECTION 07

Designing for behaviour, not recall

Seven principles that follow from the evidence rather than from convention.

Each principle below is derived from a specific finding in Sections 2 and 5. None requires a particular product; all require a decision about how the programme is structured and who owns it.

01 Set cadence from the decay curve, not the compliance calendar

Detection skill returns to baseline at roughly six months and is restored by a reminder.[17] The researchers' own recommendation is reinforcement at approximately four-month intervals. That implies three to four touchpoints a year, most of them short. It does not imply three to four full-length annual modules — the reminder formats that worked in the study were brief videos and interactive examples.

02 Target roles and measured susceptibility, not headcount

Risk concentrates: 8% of employees drive the majority of insecure data handling,[2] 30.6% of clickers click again,[16] and two-thirds of social-engineering cases target privileged accounts.[6] Build a high-risk cohort from entitlements plus observed behaviour, and give it materially more attention than the general population. Help desk, finance, IT administration, HR and executive assistants belong in it by default.

03 Train the verification procedure, not the warning signs

AI has degraded the surface cues that "spot the phish" training depends on, while under 2.5% of AI-assisted techniques are actually novel.[1] The durable skill is procedural: when a request involves money, credentials, access or data, confirm it through a channel the employee initiates, using contact details the employee looks up. This transfers across email, voice, video, chat and whatever arrives next, because it does not depend on recognising the lure.

04 Engineer the help desk as a control, not a courtesy

Credential-reset and MFA-reset workflows are now an initial access vector at national-retailer scale.[9,10] Treat identity verification as a scripted control with no discretionary override: a defined proof set, a callback to a directory-sourced number, a hard stop for high-entitlement accounts, and explicit authority for the agent to refuse. Rehearse it against a live pretext at least quarterly.

05 Make reporting the primary behaviour you reward

Reporting is the one employee behaviour with strong independent evidence behind it — crowdsourced reports functioned as an effective live-detection mechanism in the ETH Zurich study even as embedded training failed.[16] It also directly attacks the containment gap worth $7.7M annually in the insider-risk data.[5] One-click reporting, acknowledged within minutes, with visible thanks and no consequence for a false positive or for having clicked first.

06 Remove punishment from the design entirely

Punitive simulation suppresses the reporting of real attacks — the practitioner consensus is consistent[22] and the measured stress and self-efficacy costs are documented.[25] Repeat failure is a signal to change the intervention for that person, not to escalate consequences. The same study found 83.6% of employees rated a well-run campaign as worthwhile, so buy-in is not the thing being traded away.

07 Vary the format, and shorten it

Identical repeated stimuli habituate measurably after the second exposure; varied ones do not.[18] Three-quarters of employees spend under a minute on embedded training material.[14] Design for the minute you will actually get: one behaviour, one format the employee has not seen recently, delivered close to the moment of risk.

A maturity ladder

LEVEL	CHARACTERISTIC	PRIMARY METRIC	TYPICAL GAP
1 — Compliance	Annual module, completion tracked, generic content	Completion %	No evidence of behaviour change; fails on currency, verification and reporting gaps
2 — Simulation	Annual module plus periodic phishing simulation	Click rate	Metric dominated by lure difficulty; punitive variants suppress reporting
3 — Segmented	Role-based content, high-risk cohort, quarterly reinforcement	Report rate; cohort susceptibility	Still reactive to new techniques; third-party exposure unmanaged
4 — Behavioural	Cadence set by decay, verification procedures engineered, help desk hardened, no-fault reporting	Time-to-report; verification adherence	Requires dedicated staffing; SANS data indicates ~2.8 FTE to influence behaviour[35]
5 — Managed	Human risk treated as a control domain with owners, telemetry, supplier requirements and board reporting	Containment rate; measured loss avoidance	SANS associates 4+ FTE with organisational culture change[35]

The SANS 2025 Security Awareness Report, surveying more than 2,700 practitioners across 70+ countries, found 80% ranking social engineering as their top human risk and identified staffing thresholds of roughly 2.8 FTE to influence behaviour and 4+ FTE to shift culture. This is practitioner survey data, not a controlled study, and should be read as a benchmark rather than a causal finding.[35]

SECTION 08

Measuring what matters

Replacing metrics that can be gamed with metrics that predict outcomes.

A measurement set is a statement about what a programme is for. Completion rate says it exists to satisfy an auditor. Click rate says it exists to reduce a number the programme itself controls, since the programme chooses the lures. The set below is built around behaviours the evidence shows to be consequential.

METRIC	DEFINITION	WHY IT BELONGS
Report rate	Reports as a proportion of delivered lures, per cohort	The only employee behaviour with strong independent evidence of operational value.[16] Not gameable by lure selection the way click rate is.
Time to first report	Median minutes from delivery of a campaign to the first employee report	Directly drives containment. The insider-risk data attaches a $7.7M annual differential to fast versus slow containment.[5]
Report-to-click ratio	Reports divided by clicks for the same campaign	Captures whether the workforce's net effect on a campaign is defensive. A ratio above 1 means more people helped than hurt.
Verification adherence	Share of high-risk requests (credential reset, payment change, entitlement grant) completed with documented out-of-band verification	Measures the procedure that stops the attacks in Section 2, rather than measuring recognition.
Repeat-susceptibility	Proportion of the workforce failing two or more times in a rolling 12 months	30.6% of clickers click again;[16] this identifies the cohort where intervention pays.
Guidance latency	Days from a technique appearing in threat intelligence to corresponding guidance reaching affected staff	The direct measure of the currency gap. Shadow AI showed what a twelve-month latency costs.
High-risk coverage	Proportion of privileged and finance-authorising roles with current role-specific training	Two-thirds of social-engineering cases target privileged accounts.[6] Organisation-wide averages hide this.
Sanctioned-path usage	Proportion of AI use, file sharing and remote access through approved tooling	Proxy for the friction gap. 67% of AI use runs through non-corporate accounts.[1]

What to retire, and how to say so

Completion rate should be retained for audit and removed from the risk dashboard; it measures administration. **Click rate** should be demoted from headline to diagnostic — the Purdue reproduction showed it moving from 7.0% to 15.0% on lure difficulty alone with no significant training effect,[15] so a falling click rate reported without holding difficulty constant is not a defensible claim of improvement. If it is reported, report the Phish Scale difficulty alongside it. **Vendor phish-prone benchmarks** are useful for orientation and unsuitable as a target, for the reasons in Section 5.5.[24]

REPORTING THIS TO A BOARD

Boards respond badly to metrics that only ever improve, because they correctly infer the metric is being managed rather than the risk. A credible human-risk report contains: exposure sized as a range (Section 4.3); two or three behavioural metrics with direction and confidence interval; guidance latency as a process measure; and an explicit statement of what the programme cannot yet demonstrate. Saying "we cannot causally attribute loss reduction to training, and here is the independent evidence on why that is hard" is a stronger position than a number that will not survive the first informed question.

SECTION 09

A twelve-month roadmap

Sequenced so that each quarter produces evidence the next quarter depends on.

QUARTER	FOCUS	ACTIONS	EVIDENCE PRODUCED
Q1 **Establish**	Baseline and segment. Stop measuring what you cannot act on.	Build the high-risk cohort from entitlements and incident history. Baseline report rate, time-to-report and repeat-susceptibility. Map current programme against the Section 6 table. Audit help-desk identity verification against the actual procedure in use, not the documented one.	A segmented risk picture and a compliance gap list. First honest report-rate baseline.
Q2 **Harden**	Fix the two controls that stop the highest-cost attacks.	Rewrite and enforce help-desk identity verification with callback and hard stops for privileged accounts. Publish an out-of-band verification standard for payments, credential changes and entitlement grants, with named executive backing for refusal. Deploy one-click reporting with acknowledged response.	Verification adherence becomes measurable. Report rate should move first here.
Q3 **Re-cadence**	Move from annual delivery to decay-driven reinforcement.	Replace one annual module with short reinforcement at roughly four-month intervals, varied in format. Build role-specific content for help desk, finance, IT admin, HR and executive support. Introduce voice and multi-channel scenarios — not only email. Publish AI-use guidance and close the shadow-AI gap explicitly.	Guidance latency becomes a tracked metric. Cohort susceptibility differentials become visible.
Q4 **Extend**	Push the boundary out to suppliers and up to the board.	Add awareness-programme evidence to third-party due diligence for suppliers with production or identity access. Run a live pretext exercise against the help desk and a deepfake-scenario tabletop with finance and executive teams. Produce the first board-level human-risk report on the Section 8 set.	A defensible position on the 48% of breaches involving third parties, and a board metric set that survives challenge.

What to do in the first thirty days

1. **Call your own help desk.** Have someone the agents do not know attempt a credential reset using only information available from LinkedIn and the corporate website. That single test will tell you more than a quarter of simulation data.
2. **Measure report rate and time-to-first-report** on your next campaign and put them on the same slide as click rate. If report rate is not instrumented, that is itself the finding.
3. **Ask when your AI guidance was published** relative to when AI use reached half your workforce. That difference is your guidance latency, and probably your largest current gap.
4. **Check whether your simulation programme has consequences attached.** If failing twice notifies a manager, you have a reporting-suppression mechanism working against the behaviour with the best evidence behind it.

CONCLUSION

Treat human risk as a control domain, not a training obligation

The evidence assembled here supports an uncomfortable but useful conclusion: the industry has been buying the wrong thing, measuring it with the wrong metric, and delivering it on the wrong schedule — and the breach data has been saying so, consistently, for five years.

The 62% figure has not moved because the dominant intervention is aimed at only one of seven distinct failure modes, and does not survive six months even against that one. Meanwhile the attack has moved to channels most curricula do not cover, aimed at roles most programmes do not segment, using tools that remove the tells most content still teaches.

The path forward is not more training. It is a different structure: cadence set by measured decay rather than by the compliance calendar; targeting set by entitlement and observed behaviour rather than by headcount; content built around a verification procedure that transfers across channels rather than around recognition cues that AI has obsoleted; a help desk engineered as a control; reporting treated as the primary behaviour and never punished; and a measurement set a board can interrogate without the answers falling apart.

None of that is exotic, and most of it is cheaper than what it replaces. What it requires is a willingness to separate the compliance obligation from the risk-reduction objective, fund them differently, and stop claiming that the first accomplishes the second.

Compliance answers the question "did we train them?" Risk answers the question "would they verify?" Only one of those has ever stopped a breach.

Safe Instinct

SafeInstinct builds cybersecurity awareness training that turns security and privacy requirements into practical habits teams can use the same day. Our work is grounded in the same research reviewed in this paper: short, varied reinforcement rather than annual modules; role-specific content for the people whose access actually carries risk; verification behaviour rather than recognition cues; and reporting treated as the outcome worth building for.

If you would like to discuss how this evidence applies to your programme, or want a copy of the underlying source list, get in touch at safeinstinct.com.

REFERENCES

Sources

1 Verizon. *2026 Data Breach Investigations Report.* May 2026. verizon.com/business/resources/reports/dbir/

2 Verizon. *2025 Data Breach Investigations Report* and Executive Summary. April 2025. verizon.com/business/resources/reports/2025-dbir-executive-summary.pdf

3 IBM Security and Ponemon Institute. *Cost of a Data Breach Report 2026.* 29 July 2026. 602 organisations; study period March 2025 – February 2026. ibm.com/reports/data-breach

4 Federal Bureau of Investigation, Internet Crime Complaint Center. *2025 Internet Crime Report.* 2026. ic3.gov/AnnualReport/Reports/2025_IC3Report.pdf

5 Ponemon Institute and DTEX Systems. *2026 Cost of Insider Risks: Global Report.* May 2026. 354 organisations; 7,490 incidents. ponemon.dtex.ai

6 Palo Alto Networks Unit 42. *2025 Global Incident Response Report: Social Engineering Edition.* Casework May 2024 – May 2025. unit42.paloaltonetworks.com/2025-unit-42-global-incident-response-report-social-engineering-edition/

7 Mandiant / Google Cloud. *M-Trends 2026.* March 2026. cloud.google.com/blog/topics/threat-intelligence/m-trends-2026

8 CrowdStrike. *2025 Global Threat Report.* February 2025. Vendor telemetry. crowdstrike.com/en-us/press-releases/crowdstrike-releases-2025-global-threat-report/

9 UK National Cyber Security Centre, reported commentary on the 2025 UK retail intrusions (M&S, Co-op, Harrods), including SIM-swap and help-desk impersonation tradecraft. 2025.

10 Marks and Spencer Group plc. Half-year results for the 26 weeks ended 27 September 2025, published 6 November 2025.

11 National Cybersecurity Alliance and CybSafe. *Oh Behave! The Annual Cybersecurity Attitudes and Behaviors Report 2025–2026.* 30 September 2025. 6,500+ respondents across seven countries. staysafeonline.org/press/study-65-now-use-ai-but-majority-remain-untrained-on-risks

12 Resilience. *H1 2026 Claims Report*, as reported by Claims Journal, 30 July 2026. claimsjournal.com/news/national/2026/07/30/339184.htm

13 Coalition. *2025 Cyber Claims Report.* coalitioninc.com/blog/cyber-insurance/2025-cyber-claims-report

14 G. Ho, A. Mirian, E. Luo, K. Tong, E. Lee, L. Liu, C. A. Longhurst, C. Dameff, S. Savage, G. M. Voelker. "Understanding the Efficacy of Phishing Training in Practice." *46th IEEE Symposium on Security and Privacy*, 2025. ~19,500 employees; 8 months; 10 campaigns.

15 A. T. Rozema and J. C. Davis. "Anti-Phishing Training (Still) Does Not Work: A Reproduction of Phishing Training Inefficacy Grounded in the NIST Phish Scale." arXiv:2506.19899, October 2025. 12,511 employees. arxiv.org/abs/2506.19899

16 D. Lain, K. Kostiainen, S. Čapkun. "Phishing in Organizations: Findings from a Large-Scale and Long-Term Study." *IEEE Symposium on Security and Privacy*, 2022. 14,733 employees; 15 months; 117,864 emails. arxiv.org/abs/2112.07498

17 B. Reinheimer, L. Aldag, P. Mayer, M. Mossano, R. Duezguen, B. Lofthouse, T. von Landesberger, M. Volkamer. "An investigation of phishing awareness and education over time: When and how to best remind users." *USENIX SOUPS 2020.* 409 participants; 12 months. usenix.org/system/files/soups2020-reinheimer_0.pdf

18 B. B. Anderson, C. B. Kirwan, J. L. Jenkins, D. Eargle, S. Howard, A. Vance. "How Polymorphic Warnings Reduce Habituation in the Brain — Insights from an fMRI Study." *ACM CHI 2015.*

19 B. Stanton, M. F. Theofanos, S. Spickard Prettyman, S. Furman. "Security Fatigue." *IT Professional* 18(5), NIST, 2016. csrc.nist.gov/pubs/journal/2016/09/security-fatigue/final

20 J. Prümmer, T. van Steen, B. van den Berg. "A systematic review of current cybersecurity training methods." *Computers & Security* 136, art. 103585, 2023. 142 papers reviewed.

21 G. Petrič and J. N. Just. "Information security culture and phishing-reporting model: structural equivalence across Germany, UK and USA." *Journal of Cybersecurity* 11(1), Oxford, 2025.

22 B. Krebs. "Should Failing Phish Tests Be a Fireable Offense?" *Krebs on Security*, 2019 — practitioner commentary from PhishLabs and Cofense on reporting suppression. See also J. Budge et al., "Rotten Phish Spoils Employee Experience," Forrester, 2020.

23 R. Toth, R. A. Dubniczky, O. Limonova, N. Tihanyi. "Sustaining Cyber Awareness: The Long-Term Impact of Continuous Phishing Training and Emotional Triggers." *IEEE BigData 2025.* 20 organisations; 1,300+ employees; 12 months. arxiv.org/html/2510.27298v1

24 KnowBe4. *2026 Phishing by Industry Benchmarking Report.* 42 million simulated exercises; 14.8 million employees; 64,000 organisations. Vendor-produced and vendor-analysed; no untrained control group.

25 M. Schöps, M. Gutfleisch, E. Wolter, M. A. Sasse. "Simulated Stress: A Case Study of the Effects of a Simulated Phishing Campaign on Employees' Perception, Stress and Self-Efficacy." *33rd USENIX Security Symposium*, 2024. 16,000+ employee organisation; 408 questionnaire responses.

26 US Department of Health and Human Services, Office for Civil Rights. Settlement with BST & Co. CPAs, LLP, $175,000 with corrective action plan, August 2025.

27 Google Threat Intelligence Group. Reporting on UNC6040 voice-phishing campaigns against Salesforce environments, June 2025, and subsequent disclosures, August 2025 – January 2026.

28 CrowdStrike. *2026 Global Threat Report.* Vendor telemetry. crowdstrike.com/en-us/blog/crowdstrike-2026-global-threat-report-findings/

29 CNN. "Arup revealed as victim of $25 million deepfake scam," 16 May 2024. Incident occurred January 2024 in Hong Kong.

30 Anthropic. "Disrupting the first reported AI-orchestrated cyber espionage campaign," 13 November 2025. Vendor disclosure regarding abuse of its own platform. anthropic.com/news/disrupting-AI-espionage

31 Netskope Threat Labs. *Cloud and Threat Report 2026.* January 2026. Vendor telemetry.

32 Reporting on the Salesloft Drift OAuth token compromise (UNC6395), August–September 2025, including Google threat-intelligence advisories and FBI indicators of compromise released 12 September 2025.

33 Microsoft. *Microsoft Digital Defense Report 2025*, and "Think before you ClickFix," Microsoft Security Blog, 21 August 2025.

34 Regulation (EU) 2022/2554 (DORA) Art. 13(6); Directive (EU) 2022/2555 (NIS2) Arts. 20 and 21; Regulation (EU) 2024/1689 (AI Act) Art. 4; PCI DSS v4.0.1 Reqs. 5.4.1 and 12.6; 23 NYCRR 500.14(a)(3); 45 CFR 164.308(a)(5); ISO/IEC 27001:2022 Annex A 6.3 and Cls. 7.2–7.3; NIST CSF 2.0 PR.AT and NIST SP 800-50r1 (2024); GDPR Arts. 32 and 39(1)(b); SEC Reg. S-K Item 106 and Form 8-K Item 1.05.

35 SANS Institute. *2025 Security Awareness Report* (10th anniversary edition), August 2025. Survey of 2,700+ practitioners across 70+ countries. Practitioner survey data, not a controlled study.

36 UK Information Commissioner's Office. Monetary penalty notice, Interserve Group Ltd, £4.4 million, October 2022.