The Welcome End of an Era

It’s Finally Time to Retire Comparative Sufficiency

David Billeter - August 2026

Abstract

For the first time in the nineteen-year history of the Verizon Data Breach Investigations Report, the exploitation of software vulnerabilities is the leading initial access vector in confirmed breaches, while the proportion of known exploited vulnerabilities remediated has fallen and the median time to resolution has risen. This paper argues that the significance of artificial intelligence for enterprise security is not primarily that it has made adversaries faster, but that it has invalidated the assumptions encoded in defensive tooling: that exploit development requires scarce expertise, that convincing social engineering does not scale, that business processes are self-authenticating, and that vulnerability discovery is bounded by human research capacity. Drawing on frontier-model capability research and the 2026 breach record, it examines the inversion by which discovery has outrun remediation and the two compromises through which enterprise defense became partial by design. It concludes that comparative sufficiency, the practice of judging defensive adequacy against peers rather than against the adversary, has ceased to be viable, because the displacement of attacker effort on which it depended is no longer available.

Introduction

In May 2026, Verizon published the nineteenth edition of its Data Breach Investigations Report, and for the first time in the series' history, the exploitation of software vulnerabilities was the leading initial access vector in confirmed breaches. It accounted for 31 percent of breaches, up from 20 percent the prior year, a 55 percent increase in a single reporting cycle. Credential abuse, which had led the previous edition and which most enterprise security programs have spent a decade organizing themselves around, fell to 13 percent [1].

The same report showed the defensive side moving in the opposite direction. Only 26 percent of vulnerabilities listed in CISA's Known Exploited Vulnerabilities catalog were fully remediated during the period, down from 38 percent the year before. The median time to full resolution rose from 32 days to 43. In the median case, organizations faced 50 percent more critical vulnerabilities than the prior year, and the raw count of open vulnerability instances climbed roughly eightfold. Exploitation accelerated, remediation decelerated, and the volume requiring remediation grew.

The obvious explanation is artificial intelligence, and the obvious explanation is wrong, or at least premature. The data underlying that report covers incidents between November 2024 and October 2025, which predates the arrival of the frontier models now capable of autonomous vulnerability discovery at scale. Whatever caused the exploitation surge and the remediation collapse, it was not AI-assisted vulnerability research, because that capability had not yet been deployed at the volume that would show up in this data. The traditional model of enterprise vulnerability management was already failing on its own terms before the technology that will stress it further arrived.

That sequence points toward a thesis somewhat different from the one usually offered. The common framing holds that AI has made attackers more capable and that defenders must therefore run faster. That framing is not false, but it is shallow. It treats AI as a quantitative change in adversary speed, when what the evidence actually shows is a qualitative change in what defensive tooling can be relied upon to do. Nearly every major control in the enterprise security stack encodes an assumption about attacker cost, attacker skill, or attacker speed. Severity ratings assume that exploit development is hard. Patch cadences assume that weaponizing a disclosed vulnerability takes expert-weeks. Email filtering assumes that mass-produced deception looks mass-produced. Fraud controls assume that an outsider cannot convincingly participate in the organization's own business processes. Each of those assumptions was reasonable when it was encoded. Each is now being invalidated, and the tools built on them do not fail loudly when their premises expire. They continue operating, producing outputs that look correct, while the thing they were measuring quietly stops being the thing that matters.

A metaphor that circulates in the security community in several variants captures how the operating logic of enterprise defense has shifted under this pressure. The cast changes with the telling: zebras and hyenas in one version, zebras or antelope and lions in others, two campers and a bear in the joke form most people encounter first [2]. The structure never changes. The prey does not need to be the fastest of its kind, only not the slowest, because the predator's appetite will be satisfied before the faster animals are reached.

Applied to cybersecurity, the metaphor described a market for adversary attention: the attackers would exhaust their capacity on the weakest targets, and any organization that had invested enough to look harder than its peers would be passed over. The resulting posture is one of comparative sufficiency, meaning a defensive standard judged adequate by reference to peers rather than by reference to the adversary. It has been employed in executive suites and embedded in how programs are funded, benchmarked, and defended to boards. It is worth being precise about what comparative sufficiency actually purchases, because the metaphor is polite about it. What relative hardness buys is not protection but displacement: the adversary's effort moves to a softer target. The strategy succeeds only when someone else is breached instead, and it was tenable only while adversary capacity was scarce enough for displacement to be real. It assumed a bounded population of capable attackers and a bounded capacity for parallel operation. Both boundaries have moved, and the paper

returns to this framing in its conclusion, because the criterion for survival in an environment where the predators have industrialized is not comparative sufficiency. It is absolute adequacy against the adversary, and absolute adequacy is a completeness standard that partial defensive postures do not meet.

What follows examines the breach record as it now stands, the adversary economy that industrialized before AI arrived and was therefore positioned to absorb it immediately, the four assumptions that AI has invalidated and the tools built on each, the inversion in which vulnerability discovery has outrun the human capacity to remediate, and the two compromises through which enterprise security became partial by design. The argument is not that defense is hopeless. It is that defensive postures must now become comprehensive, and that acceptance of partial protection is no longer an option.

What the Breach Record Now Shows

The 2026 Data Breach Investigations Report analyzed more than 22,000 confirmed breaches across 145 countries, drawn from incidents occurring between November 2024 and October 2025. Its headline finding is the reordering of initial access vectors. Exploitation of vulnerabilities reached 31 percent, its first appearance at the top of the ranking in the report's nineteen-year history. Phishing accounted for 16 percent, credential abuse for 13 percent, and pretexting for 6 percent. Beyond the initial access vectors, ransomware appeared in 48 percent of breaches, up from 44 percent. The human element remained present, appearing in 62 percent of breaches [1].

The reordering is less dramatic than it first appears. Verizon tracks identity-related initial access across three separate categories, and phishing, credential abuse, and pretexting together account for 35 percent, which exceeds the vulnerability exploitation figure. The headline comparison of 31 percent against 13 percent sets a consolidated category against a fragmented one. Identity-based access has not receded; it has been outpaced by a category that grew faster [3].

What is not ambiguous is the remediation picture. The proportion of KEV-listed vulnerabilities fully remediated fell to 26 percent from 38 percent. Median time to full resolution rose to 43 days from 32, an increase of roughly a third. Organizations in the median case confronted 50 percent more critical vulnerabilities than in the prior period. Open vulnerability instances rose eightfold [1]. These are not the numbers of a discipline that is holding steady under pressure. They describe a function that is being overwhelmed, and they describe it in a period before the AI capabilities that will overwhelm it further had been deployed.

The report's treatment of artificial intelligence is more restrained than most industry commentary, and more useful for it. Verizon collaborated with Anthropic's Safeguards Team to analyze 793 threat actors who received enforcement action for violating acceptable use policy between March 2025 and February 2026. In the median case, an actor sought AI assistance across roughly 15 distinct MITRE ATT&CK techniques.

Regarding AI-assisted initial access specifically, 44 percent mapped to phishing and 32 percent to exploiting vulnerabilities. Fewer than 2.5 percent of the techniques observed were classified as rare [3], [4].

That last figure is the important one, and it cuts against the more excitable readings of AI in offensive security. The picture it describes is not one of adversaries inventing novel attack classes. It is one of adversaries applying AI to the techniques they already used, across more of the attack chain, at greater speed and volume. AI is currently functioning as an operational multiplier on known tradecraft rather than as a generator of unknown tradecraft. This is consistent with what the rest of the breach record shows, and it should temper any claim that the threat landscape has been transformed into something unrecognizable.

It should not, however, be read as reassurance. If AI were producing novel attack techniques, the defensive response would be to develop detection for those techniques. What the data shows is AI being applied to techniques the industry has known about for decades and has repeatedly failed to close. Multiplying the throughput of a known attack against a control that was already only partially deployed does not create a new problem. It converts a tolerated problem into an untenable one, which is harder to fix, because the remedy is not a new detection capability but the completion of work that was deliberately left incomplete. In many cases comparative sufficiency is what made those controls tolerable in the first place: they were left incomplete deliberately, because peer comparison suggested incompleteness was survivable.

The Industrialized Adversary

The adversary economy that received these AI capabilities was not a loose community of individual operators. It was an industry, already organized and relatively efficient, and its structure explains why AI capability was absorbed so quickly and distributed so widely. Understanding that structure is necessary before assessing what AI changed, because the answer depends heavily on what was already in place.

The defining structural development of the past decade is the separation of the attack lifecycle into discrete, tradeable functions performed by specialists who need not know one another, trust one another beyond the terms of a transaction, or share any objective beyond completing an exchange. The clearest expression is the ransomware-as-a-service affiliate model. A developer team builds and maintains the platform, comprising the encryption tooling, payment and decryption infrastructure, negotiation portal, and leak site, but conducts no intrusions. Affiliates conduct the intrusions and deploy the payload, typically retaining sixty to eighty percent of the ransom. LockBit, the most active group by victim count from 2022 through early 2024, operated on this model with 194 affiliates documented in records seized during Operation Cronos in February 2024 [5].

This observation has a scholarly pedigree that predates the current threat landscape by a decade, and it is worth acknowledging directly, because it clarifies what is and is not

new. In 2015, Kurt Thomas and nine co-authors spanning academia and industry, publishing at the Workshop on the Economics of Information Security, systematized the research community's understanding of the underground economy, describing internet crime as already dependent on “a loose federation of specialists” selling capabilities, services, and resources purpose-built for abuse, from which criminal entrepreneurs could assemble complete operations out of individually purchased components [6]. The same work proposed a taxonomy that remains the most useful frame for reading this market: profit centers, the activities that extract wealth from victims, and support centers, the suppliers of tooling, infrastructure, and services on which the profit centers depend. The sociologist Jonathan Lusthaus, approaching the same phenomenon through more than two hundred interviews conducted across the world's cybercrime hotspots, reached a converging conclusion in his 2018 study titled “Industry of Anonymity”, showing how cybercrime had matured into an organized, profit-driven, globally interconnected industry, long past the era of the lone actor [7], [8].

The separation of development from deployment created a secondary market for access. If the affiliate's comparative advantage is intrusion rather than tooling construction, the affiliate's bottleneck becomes the cost and time of obtaining initial access. The market's answer was the initial access broker, an actor who specializes exclusively in obtaining, documenting, and selling authenticated access without deploying any payload. Broker activity roughly doubled between 2023 and 2024, and the pricing tells the story of a maturing market: the average price of a verified network access listing fell from approximately $3,066 to $1,295, a decline of about 60 percent, with 65 percent of listings priced under $2,000 [9]. Supply expanded faster than demand.

Beneath the brokers sits the credential supply chain. Threat intelligence research covering 2024 documented the compromise of approximately 3.9 billion credentials from 4.3 million infected devices, with the top three malware-as-a-service families, Lumma, StealC, and RedLine, together responsible for more than 75 percent of infections [10]. Each operates as a subscription service through Telegram channels, priced from roughly $250 monthly for entry tiers to $1,000 for premium features such as session cookie restoration. The share of infostealer infections exposing enterprise single sign-on credentials rose from 6 percent at the start of 2024 to 16 percent in 2025, and the median interval between a personal device infection and that credential's appearance in an enterprise breach compressed to roughly seven days [11]. The Snowflake campaign of 2024, in which UNC5537 accessed the cloud environments of approximately 165 organizations, illustrates the chain in operation [12]. The credentials were not stolen by the actor who used them. They were harvested by commodity malware, aggregated through dark-web channels, and acquired in a market transaction. The distance between credential theft and enterprise breach was not technical. It was a price.

What this economy has built, under years of law enforcement pressure, is not merely capability but resilience. Operation Cronos, coordinated across ten countries in February 2024, seized 34 servers, closed 14,000 affiliate accounts, froze 200 cryptocurrency wallets, and recovered more than 1,000 decryption keys, and the UK's

National Crime Agency repurposed LockBit's own leak site to publish the operation's details [5]. Within days, LockBit's administrator had posted from a reconstituted site and resumed recruiting. Within weeks, new victims were appearing. Chainalysis recorded ransomware payments falling from $1.25 billion in 2023 to approximately $814 million in 2024, the first annual decline since 2022 [13], but the number of distinct data leak sites grew rather than shrank, with 56 new sites appearing in 2024, more than double the prior year [14]. Affiliates migrated to RansomHub, BlackSuit, and Akira. The market redistributed share, but it did not reduce volume.

That resilience is structural rather than accidental. The platform developer, the affiliate, the access broker, the infostealer operator, and the laundering service are separate entities with no organizational dependency beyond discrete transactions. Dismantling one node eliminates that node and not the market relationship it served, because other nodes exist to fill it. ALPHV exited under FBI pressure and affiliates moved on. RedLine was disrupted through Operation Magnus in October 2024, and Lumma absorbed the displaced volume, with detections rising 369 percent between the first and second halves of that year [15].

This is the system that has received AI capability, and its structure determined the reception. A modular market with low barriers to entry, subscription distribution already in place, and a population of participants selected for willingness to adopt whatever tool improves margin does not deliberate about new capability. It prices it and resells it. When the underground began packaging malicious language models, the distribution infrastructure already existed. WormGPT surfaced in mid-2023 as a stripped-down derivative trained on malware-related data, and it was followed within weeks by imitators, then shut down under pressure that August. Variants built on Mixtral and Grok appeared on BreachForums between October 2024 and February 2025, priced from roughly $110 monthly to $5,400 for private versions, and a fourth iteration surfaced in September 2025 [16]. One Cambodia-based marketplace later sanctioned by the U.S. Treasury recorded year-over-year growth in AI service vendor revenue of approximately 1,900 percent during 2024 [17]. None of this required the adversary economy to reorganize. It required only that a new category of product be listed.

Four Assumptions Artificial Intelligence Has Invalidated

Security controls are not neutral instruments. Each one encodes a model of the adversary, and that model contains estimates of what an attack costs, what skill it requires, and how long it takes. Those estimates are rarely stated explicitly and almost never revisited, because they are usually correct for long enough that nobody notices they are assumptions at all. They become the part of the background against which tools are configured, budgets are argued, and risk is accepted.

The difficulty with an expired assumption is that it produces no error message. The control continues to function, its outputs continue to look correct, and the organization continues to rely on them, while the relationship between the output and the reality it was meant to describe quietly dissolves.

That Exploit Development Requires Scarce Expertise

The most consequential assumption in enterprise vulnerability management is that turning a disclosed vulnerability into a working exploit is difficult, slow, and performed by a small population of specialists. Nearly every operational practice in the discipline depends on it. Patch cadences that release monthly, staged rollouts that reach full deployment over weeks, testing cycles that precede production deployment, and the entire tolerance for a patch gap between disclosure and remediation all assume that defenders have time because attackers need time.

That assumption was well founded for most of the field's history. Historically, exploiting a disclosed vulnerability meant patch diffing (comparing pre-patch and post-patch code or binaries to locate the change and reverse-engineer the flaw it addressed) which was slow and specialized work. The incidents that shaped defensive doctrine reflect this. WannaCry arrived 59 days after the corresponding Microsoft bulletin. The public exploit for Citrix Bleed took roughly two weeks. In Mandiant's 2020 analysis of N-day exploitation, 16 of 25 vulnerabilities took a month or more to exploit [18].

Anthropic's research on N-day exploitation, published in June 2026, shows what has happened to that interval. Testing frontier models against 18 security patches for Firefox's JavaScript engine, the strongest model produced working proof-of-concept crashes for 14 of them, including eight complete working exploits. Progression across model generations is the striking part: from Claude Opus 4.5 to Opus 4.8, the number of patches convertible to a working proof of concept rose from 2 to 11, and the Claude Mythos Preview reached 14 [18]. The first proof of concept arrived in about 12 minutes, and 13 of them within 40 minutes.

The timing comparison is the finding that should reorganize defensive thinking. Mythos Preview produced its first working exploit within an hour of Mozilla issuing the corresponding patch. The patched version of Firefox containing that fix would not be released for another 18 days. The exploit existed before the remedy was available to users, against a vendor whose patch gap is among the industry's fastest, on software that updates itself automatically and requires only a browser restart to adopt the fix. Anthropic's contention is that if patch gaps this narrow are wide enough to exploit, then most other software's gaps are certainly wide enough [18].

The Windows results extend the finding to closed-source software, where the model works from compiled binaries and decompiler output stripped of variable names, types, and structure. Across 21 Windows kernel privilege-escalation vulnerabilities, Mythos Preview produced proofs of concept for 18, the first within 31 minutes and all 18 within six hours at a cost of roughly $2,200 in API credits. It then produced eight distinct complete exploit chains taking a low-privilege user to full SYSTEM control, at a total cost of $15,700, averaging about $2,000 per privilege escalation. Measured against Windows Autopatch timelines, where a patch typically reaches 90 percent of enrolled devices in seven days and forced reboots occur on day 11, the model would have completed all eight exploit chains before any of the devices had received the update [18].

The implication for severity rating deserves particular attention, because severity rating is the mechanism through which remediation is prioritized and, as a later section argues, negotiated. Microsoft's advisories rated 14 of the 21 tested vulnerabilities as either “Exploitation Less Likely” or “Exploitation Unlikely.” Mythos Preview produced proofs of concept for 13 of those 14, including a full privilege escalation for one rated “Exploitation Unlikely.” Anthropic's conclusion is stated plainly: the rating system is currently calibrated to human researchers, and as models of this class become widely available, that calibration may need to change [18].

This is what an expired assumption looks like in its purest form. The rating did not malfunction. It correctly described the probability of exploitation by the population of attackers it was designed to model. However, that population is no longer the relevant one. Every downstream decision that treated “Exploitation Unlikely” as a reason to defer remediation inherited an estimate that was accurate when originally designed but is no longer accurate now, and nothing in the output of the rating system signals the difference.

Anthropic's summary of the operational consequence is that the typical patching playbook, with monthly release cadences, multi-week staged rollouts, and lag between pre-release and stable channels, no longer holds, because it was built on the assumption that weaponizing a patch takes expert-weeks and that the pool of capable experts is limited. Their phrasing is that “N-day” has become dangerously misleading, and that “N-hour” is closer to the operating reality. The binding constraint on N-day exploitation is now a few thousand dollars and API access rather than scarce reverse-engineering expertise [18].

That Convincing Social Engineering Does Not Scale

The second assumption governs the defense of the human attack surface. It holds that deception quality and deception volume trade off against each other. A convincing, personalized message required research into a specific target, and that research cost time, which limited the technique to targets whose value justified the investment. Mass-produced deception, conversely, carried the artifacts of mass production: generic salutations, grammatical patterns inconsistent with native composition, implausible urgency, imprecise impersonation of organizational voice and formatting.

Enterprise defense was built on that tradeoff in two layers. Technical filtering learned the artifacts of mass production and scored against them. Security awareness training taught employees to recognize the same artifacts, which is why a generation of training material emphasized checking for spelling errors and awkward phrasing. Both layers were effective in proportion to how reliably the artifacts appeared, and they appeared reliably because producing content without them at volume was expensive.

Large language models have removed the tradeoff. Controlled studies published in 2024 measured AI-generated phishing achieving a 54 percent click-through rate against approximately 12 percent for traditional phishing, roughly a 4.5-fold increase in effectiveness at near-zero marginal production cost. A University of Oxford study of

similar design found AI-generated messages producing a 60 percent higher click rate than human-written controls. IBM X-Force research found the time to produce a convincing phishing message falling from approximately 16 hours of manual work to roughly five minutes. Analysis of phishing samples collected during 2024 found AI involvement in 73.8 percent of messages examined, rising above 90 percent for samples exhibiting polymorphic or advanced evasion behavior [19], [20].

The scale implication matters more than the quality implication, and industry commentary usually reverses the emphasis. The concern most often articulated is that individual messages have become more convincing. The more consequential change is that highly convincing, individually personalized messages have become producible at volume. An adversary who can generate contextually personalized content at the cost of a prompt can now conduct against an organization's entire user population the kind of campaign that previously would have been reserved for a handful of executives, with each message tailored to its recipient's role, manager, recent project activity, and vendor relationships. The economics of the attack changed more fundamentally than the technique.

Two observations complicate the picture. The first is the DBIR finding already noted: despite 44 percent of AI-assisted initial access mapping to phishing, phishing's share of the incident dataset moved little year over year. The second is that vishing, voice-based phishing, surged 442 percent between the first and second halves of 2024, with the help desk emerging as a primary entry point [21]. Read together, these suggest that improved email filtering may be absorbing some of the increased quality while pushing adversary effort toward channels where the controls are weaker and the verification is human. That is not defensive failure. It is defensive displacement, and it means the pressure arrives somewhere the organization is usually monitoring and measuring less carefully.

That Business Processes Are Self-Authenticating

The third assumption is the oldest, the least examined, and the one least likely to appear in any control inventory, because the controls that depend on it were rarely designed as controls at all. It holds that an organization’s ordinary processes are self-authenticating: that a request arriving through the normal channel, in the normal format, from an apparently normal participant, and conforming to the normal pattern of such requests, is thereby probably legitimate. A significant funds transfer required a conversation with the approving executive because that was how the organization made financial decisions, not because anyone had threat-modeled the conversation as an anti-fraud mechanism. A change to pricing, shipping instructions, or vendor banking details required a ticket, an approval, and a system entry, and the sequence itself was treated as evidence that the change was authorized.

The assumption was reasonable because it rested on something real: participating convincingly in an organization’s internal processes was genuinely difficult for an outsider. It required knowing which system holds the authoritative record, who approves which category of change, what a routine request looks like in tone and format, which

thresholds trigger additional scrutiny, and when in the cycle such requests ordinarily arrive. That knowledge was tacit, distributed across employees, and expensive to acquire from outside. The process was the control, and its security value derived almost entirely from the outsider's ignorance of it.

Awareness training reinforced the same logic from the other direction. The instruction to treat requests falling outside normal process as suspicious presumes that an attacker will operate outside the normal process, because the normal process is the thing they cannot convincingly enter. Nearly every behavioral control an enterprise deploys against fraud and social engineering inherits that presumption.

What has collapsed is the cost of acquiring this knowledge. Extended dwell time inside a compromised environment gives an adversary access to the ticketing system, the approval chains, the email archives, and the internal documentation, which together constitute a working manual for how the organization authorizes things. Large language models make sense of that material at a speed and scale that manual review never permitted, and they generate requests indistinguishable in register and format from the genuine article. The result is an attack that does not involve stolen credentials being used in an unusual way, or malware, or any deviation from procedure. It involves a correctly formatted request, submitted through the proper channel, approved by the right person, and executed by a system that has no basis for refusing it. In a strange way, an adversary may document the actual processes of an organization far better than those processes are understood in the organization itself.

The consequences are not confined to fraudulent payments, which is the variant most organizations have modeled. An adversary who understands the process can have the retail price of an item changed, so that goods are purchased legitimately at a fraction of their value. They can have shipping instructions redirected to an address that has been added, through the standard vendor onboarding workflow, to the list of approved destinations. They can have a supplier's banking details updated, a user provisioned into a privileged group, an approval threshold raised, a security control excepted for a named system, or a scheduled reconciliation suppressed. Each of these is an ordinary business change that some employee is authorized to make and some system is designed to execute. None requires the adversary to break anything. The organization performs the attack on itself, correctly, through the mechanisms built for the purpose.

This is also why such attacks are detected late when they are detected at all. Security monitoring is oriented toward technical anomaly, and there is no technical anomaly to find. The authentication succeeded, the authorization was valid, the workflow completed, and the audit log shows a properly approved change. Discovery typically comes from the business rather than from security, when the margin on a product line is inexplicable, when goods arrive somewhere unexpected, or when a supplier reports non-payment, and by then the transaction has been complete for weeks and is indistinguishable in the record from thousands of legitimate ones.

Synthetic media is the accelerant on this fire rather than the fire itself, and the distinction matters for how organizations respond. What deepfakes remove is the last remaining

friction in the process attack, which is the moment when a human participant is asked to confirm that they are who they claim to be. The January 2024 incident at the engineering firm Arup illustrates the combination. A finance employee in the Hong Kong office authorized the transfer of approximately HK$200 million, roughly $25.6 million, across 15 transactions to five bank accounts, following a video conference in which AI-generated likenesses of the company's UK-based chief financial officer and other colleagues appeared alongside the real employee. The video conference was not a deviation from normal enterprise process. It was normal enterprise process, conducted with participants whose identities had been fabricated. The employee who followed the prescribed procedure correctly, joining the call, observing the participants, and completing the transactions, was doing exactly what a successful attacker required [22], [23]. Note what the deepfake actually contributed: the attack still required knowing that a transfer of that size needed a video conference with that particular executive, that such conferences were routine, and how the resulting instruction would be executed. The synthetic video satisfied the one control the process knowledge could not.

Hong Kong police, disclosing the case and related investigations, reported that AI deepfakes had been used on at least 20 occasions to defeat facial recognition tied to identity card verification, supporting 90 fraudulent loan applications and 54 bank account registrations from a single batch of stolen identity cards. Microsoft's 2025 Digital Defense Report documents rising prevalence of synthetic identities, deepfake video used against facial recognition, and AI-generated identity documents deployed against know-your-customer systems [24], [22].

Voice extends the same pattern to the help desk, which is the process most consistently targeted because it exists specifically to help people who cannot authenticate. Commercial voice cloning services at consumer price points can produce convincing simulations from a reference sample of a few minutes, and for any publicly facing executive, or any employee who has appeared on a recorded call, an earnings presentation, a conference panel, or a training video, that sample is available without any intrusion. The verification question a help desk agent asks was designed to resist impersonation by an actor who sounds different from the person being impersonated and who lacks the biographical detail. It was not designed to resist an actor who sounds identical and who has assembled the biographical answers in advance from open sources. Scattered Spider's 2023 intrusion at MGM Resorts, which began with a phone call to a help desk referencing detail harvested from LinkedIn, demonstrated the value of the channel before synthetic voice made it substantially easier to work [25]. The agent in that case followed the process. The process was the vulnerability.

The defensive implication differs from the one usually drawn. Detecting synthetic media is a useful capability and organizations should pursue it, but treating it as the answer misidentifies the problem, because it addresses only the variant of process abuse in which the adversary needs to appear as a person. The broader requirement is that consequential business changes should not be self-authorizing on process conformance alone. Verification needs to rest on something the adversary cannot obtain by observing how the organization works: out-of-band confirmation through a channel

the requester did not choose, controls that trigger on the substance of a change rather than on the correctness of its paperwork, and monitoring that treats an unusual price adjustment or a new shipping destination as a security-relevant event rather than as a business transaction that happens to be logged. This is the completeness argument again, arriving from a different direction. The gap is not in any individual control but resides in the space between the security function, which watches authentication and endpoints, and the business function, which owns the processes where the loss actually occurs.

That Vulnerability Discovery Is Bounded By Human Researchers

The fourth assumption is the broadest and the one whose invalidation is least visible in the breach record, because it is arriving now rather than having already arrived. It holds that the rate at which vulnerabilities are discovered is governed by the size and productivity of the security research community, and that this rate, while it varies, is stable enough to plan against. Every element of the vulnerability disclosure and remediation ecosystem rests on it: the 90-day coordinated disclosure convention, the staffing of vendor security response teams, the throughput assumptions of maintainer communities, and the implicit expectation that the queue of known unpatched vulnerabilities, while long, is finite and roughly steady.

Anthropic's Project Glasswing provides the clearest available measurement of what happens when that assumption fails. Within roughly one month of launch, Anthropic and approximately 50 partners had used Claude Mythos Preview to identify more than ten thousand high- or critical-severity vulnerabilities across systemically important software. Several partners reported their rate of bug-finding increasing by more than a factor of ten. Cloudflare identified 2,000 bugs across critical-path systems, 400 of them high or critical severity, at a false positive rate the team considered better than human testers. Mozilla found and fixed 271 vulnerabilities in Firefox 150 while testing the model, more than ten times the number found in the prior version using Claude Opus 4.6. The UK's AI Security Institute reported the model as the first to solve both of its multistep cyber ranges end to end [26].

Separately, Anthropic scanned more than 1,000 open-source projects and recorded 6,202 estimated high- or critical-severity findings out of 23,019 total. Of 1,752 high- or critical-rated findings independently assessed by six security research firms, 90.6 percent proved to be valid true positives and 62.4 percent were confirmed as high or critical severity, putting the effort on track for nearly 3,900 confirmed high- or critical-severity vulnerabilities in open-source code alone [26].

One further data point suggests how little sophistication is required to convert autonomous capability into real compromise. In July 2026, Anthropic disclosed that a review of 141,006 cybersecurity evaluation runs had identified three incidents in which a Claude model, operating in an evaluation environment misconfigured to permit internet access while its prompt stated it had none, reached real systems and compromised the production infrastructure of three organizations. The techniques used were not exotic: weak passwords, unauthenticated endpoints, an exposed debug page, SQL injection. In

one run the model scanned roughly 9,000 targets before finding one it could compromise. In another it published a malicious package to a public registry that was downloaded and executed on 15 real systems within roughly an hour, exfiltrating credentials from a security company's scanner. The organizations affected had not detected the activity and had not contacted Anthropic; they learned of it when Anthropic reached out. Autonomous agents did not require novel exploits to reach production systems. They required only the basics, executed at machine scale, against defenses that were incomplete [27].

The Inversion

The expected consequence of AI-assisted vulnerability discovery was that attackers would find more flaws. The observed consequence, at least so far, is that defenders have found more flaws than they can fix, and that the constraint on software security has moved from one end of the process to the other.

Anthropic states the change directly in its Project Glasswing update: progress on software security used to be limited by how quickly new vulnerabilities could be found, and is now limited by how quickly they can be verified, disclosed, and patched [26]. That sentence describes a reversal of the discipline's organizing constraint, and the operational evidence behind it is unusually candid for a vendor publication.

Of the high- or critical-severity vulnerabilities Anthropic disclosed to open-source maintainers, 530 had been reported at the time of the update and 75 had been patched, with 65 receiving public advisories. A further 827 confirmed vulnerabilities were awaiting disclosure. The average time to patch a high- or critical-severity finding was two weeks. Several maintainers reported being severely capacity constrained, and some asked Anthropic to slow its rate of disclosure because they needed more time to design patches. Anthropic acknowledges that even at a deliberately restrained pace, the effort is adding to an already overloaded security ecosystem [26].

The dynamic compounds because maintainers were already absorbing a deluge of low-quality AI-generated bug reports from other sources, which raises the cost of triaging any individual report and lowers the credibility of the queue as a whole. Anthropic's own triage process involves reproducing each finding, reassessing severity, checking for existing fixes, and writing a detailed report, performed either internally or by one of six external security research firms. That is substantial human effort applied to each item, and it is the reason the disclosure pipeline narrows so sharply at each stage.

The commercial side shows the same shape from a different angle. Following Glasswing, the most recent Palo Alto Networks release included over five times the usual number of patches. Microsoft reported that the number of new patches it releases would continue trending larger for some time. Oracle reported finding and fixing vulnerabilities multiple times faster than before [26]. These are the outputs of vendors who have adopted the capability. Each of those patches must then be tested, distributed, and applied by every downstream organization running the software, and

the volume arriving at that end of the pipeline is set by the vendors' new discovery rate rather than by the customers' absorption capacity.

Placed beside the DBIR figures, the two datasets describe one phenomenon measured from opposite ends. Anthropic reports discovery outrunning remediation at the source. Verizon reports open vulnerability instances rising eightfold and the proportion of KEV-listed vulnerabilities remediated falling from 38 percent to 26 percent at the destination [1]. The volume that maintainers and vendors cannot patch quickly enough becomes the backlog that enterprises cannot remediate quickly enough, and the interval during which a disclosed vulnerability is known but unfixed is the interval in which N-hour exploitation operates.

This does not mean AI-assisted vulnerability discovery is net harmful. Anthropic's stated purpose in Project Glasswing is to secure critical software before comparable models can be turned against it, and the wolfSSL finding alone, an exploitable certificate forgery flaw in a cryptographic library used by billions of devices, is the kind of result that justifies the effort. Nor does it mean the defensive side is losing a race, since the first substantial deployment of this capability was defensive by design. What it means is narrower and more actionable: the bottleneck has moved, and an industry whose processes, staffing, and disclosure conventions were all designed around the old bottleneck is now organized against the wrong constraint.

Two Compromises

Enterprise security is incomplete by design, and the design was not accidental. Two long-standing compromises produced it. The first is internal to the enterprise and governs how much of the known vulnerability population is actually remediated. The second is external and governs how much responsibility the producers of vulnerable software bear for the defects they ship. Both were rational adaptations to real constraints. Neither survives the current conditions and threat landscape.

The Negotiated Percentage

Vulnerability management in most large enterprises does not resolve as an engineering decision. It resolves as a negotiation between security, application, and operations teams over how much remediation work will be accepted into a delivery schedule that is already full, and is in competition with other business priorities. The negotiation produces a graduated settlement: most or all criticals, a smaller proportion of highs, a smaller proportion again of mediums, and, in practice, no lows. The security function enters that negotiation without the authority to compel and without a budget line that would fund the work independently, which means it enters from a losing position and its function becomes to lose slowly. Executive leadership is largely absent from the process, and its absence is itself a decision, because a compromise that no executive contests is a compromise that has been ratified by default.

The evidence that this is the actual operating practice rather than a cynical characterization comes from survey data that asked organizations not what they

achieved but what they would accept. Research conducted by the Ponemon Institute, sponsored by the vendor Rezilion and therefore worth reading with that interest in view, found organizations carrying an average backlog of approximately 1.1 million individual vulnerabilities, of which an average of 46 percent were remediated. The same respondents indicated their organizations would be satisfied if 29 percent of vulnerabilities in a year were remediated [28]. That figure is the compromise made explicit, stated as a target rather than confessed as a shortfall.

The supporting picture is consistent. Two thirds of organizations reported backlogs exceeding 100,000 vulnerabilities, and 54 percent said they could patch less than half of the backlog. Seventy-eight percent said high-risk vulnerabilities took longer than three weeks to patch, with 29 percent exceeding five weeks. Separate research found organizations typically remediating about 10 percent of weaknesses per month regardless of total volume [28], [29].

The compromise has always been defended on the grounds that it is risk-based, that the criticals get fixed and the rest are acceptable residual risk. The triage that sorts the population into tiers is performed using severity ratings calibrated to a threat model in which exploit development requires scarce human expertise. When a vendor rates a vulnerability “Exploitation Unlikely” and a frontier model produces a working privilege escalation for it, the sorting mechanism on which the negotiated percentage depends has been shown to misclassify precisely the items the negotiation discards. The compromise is not only that too little is fixed. It is that what gets fixed is chosen by a ruler that is bending, and the items sacrificed at the bottom of the list are sorted there by an estimate that no longer holds.

Contracting Out of Responsibility

The second compromise operates upstream. Enterprises have accepted, for four decades, that commercial software arrives containing defects, and that the producers of that software bear almost no legal responsibility for them. The mechanism is unremarkable and nearly universal: software license agreements limit liability and disclaim warranties, and the effect is that the party best positioned to prevent a defect bears the least consequence when it is exploited. James Dempsey of Stanford's Program on Geopolitics, Technology and Governance has put the mechanism plainly, observing that license agreements essentially shield companies from lawsuits through limitations of liability and disclaimers, and noting that questions about software liability have persisted for at least thirty years across several administrations [30].

The vendors do have a real argument. Some defects are genuinely unforeseeable. Threat models shift, and a design that was sound against the attacks contemplated at the time may fail against attacks developed later. Furthermore, exposure frequently depends on the customer's deployment environment, integration choices, and configuration, none of which the vendor controls. A warranty against vulnerabilities in general would be a warranty against the unknown, and no responsible vendor would offer it.

But that legitimate objection has been extended to cover territory it does not reach. In March 2024, CISA and the FBI issued a joint Secure by Design Alert directed at SQL injection, stating that the software industry has known how to eliminate these defects at scale for decades, that vulnerabilities of this kind have been considered unforgivable since at least 2007, and that CWE-89 nonetheless remained on 2023 lists of both the most dangerous and most stubborn software weaknesses. The alert was issued in response to the exploitation of SQL injection defects in a managed file transfer application that affected thousands of organizations. CISA's alert series as a whole exists to highlight widely known and documented vulnerabilities that have not been eliminated [31]. A defect class documented for over twenty years, with well-understood remedies available in standard libraries, is not unforeseeable by any construction of the word.

The distinction that survives, then, is not between vulnerabilities and no vulnerabilities. It is between defect classes that are known, documented, and remediable, for which accountability is coherent, and genuinely novel failures, for which it is not. That is the line CISA drew, and it is the line the vendors' own best argument implicitly concedes. What mechanism could enforce that line, whether procurement conditions, market-access regulation of the kind now emerging in several jurisdictions, or something else, is beyond the scope of this paper. What matters here is the consequence for the enterprise: whatever is eventually imposed on producers, the exposure created by an organization's own architecture, integration choices, and legacy decisions remains its own to remediate.

The Defender’s Position

Two objections to the foregoing deserve direct answers. The first is that defenders benefit from artificial intelligence as well, and that the asymmetry described is therefore temporary. The second is that enterprises could remediate more aggressively if they chose to, and that the negotiated percentage reflects preference rather than constraint. Both objections contain some truth, yet neither survives contact with how enterprise security actually operates.

On the first: defensive AI capability is real and improving. AI-assisted detection, anomaly analysis in authentication and session behavior, automated triage of phishing content, and AI-augmented security operations all represent genuine gains, and Anthropic's release of Claude Security in public beta, through which Claude Opus 4.7 was used to patch more than 2,100 vulnerabilities in three weeks, demonstrates that defensive application is not hypothetical. Notably, enterprises fixing their own code moved considerably faster than the open-source disclosure pipeline, precisely because they were not waiting on volunteer maintainers working through coordinated disclosure [26].

The asymmetry is not about which side has better technology. It is about deployment velocity and prerequisite conditions. An adversary acquires capability at the speed of a transaction: a tool is listed, a subscription is purchased, and the capability is operational in the same session. An enterprise acquires capability at the speed of a procurement

cycle, comprising vendor evaluation, proof of concept, contract negotiation, security review, deployment, integration with the existing stack, and tuning against the environment's specific baseline. The tool that produces security value is not the one that passes the proof of concept. It is the one that has been tuned and integrated and is operated by analysts who understand both its outputs and its failure modes, a state reached twelve to eighteen months after the decision, even in efficient implementations.

The prerequisite problem is more binding still. The most promising defensive application of AI to the access and session problem is behavioral detection, identifying anomalous session behavior, unusual authentication patterns, and activity inconsistent with established baselines. That capability requires continuous visibility across all authentication contexts in the environment, which is often a multi-year instrumentation program governed by the same budget cycle as everything else. The AI layer cannot be deployed before the foundation exists. Offensive capability carries no equivalent prerequisite.

Beneath these sits an asymmetry that receives almost no attention and may be the most consequential in practice: the attacker is permitted to fail during tuning in a way the defender is not. When an attacker tunes a campaign or a payload and it fails, the cost of failure is the attempt. There is no collateral consequence. When a defender tunes a detection system, adjusting an anomaly threshold or tightening the conditions under which access is challenged, the feedback environment is entirely different. The defender can never confirm that a configuration which has not been penetrated is correct, because the attacks it has not seen are by definition the ones its current tuning does not catch. Failure in the other direction, however, is loud. A false positive that blocks a legitimate user produces immediate escalation, a wave of support tickets, a conversation with a business unit leader, and sometimes a direct instruction to roll back the change. Security teams learn through repetition that aggressive tuning invites organizational conflict that can exceed the response to a security gap. The rational adaptation is to tune conservatively. The attacker iterates toward the threshold aggressively because overshooting costs another attempt. The defender retreats from it cautiously because overshooting costs organizational capital that the security team, not the attacker, must spend.

On the second objection, that enterprises could simply remediate more if they chose: the constraint is partly structural and partly temporal, and both parts are real. The staffing dimension is the more durable. The ISC2 Cybersecurity Workforce Study reported a global shortage approaching 4.8 million professionals as of 2024 [32]. The 2025 study, surveying 16,029 practitioners, declined for the first time to publish a single aggregate gap figure, citing instead the rise of skills shortages, particularly in AI and cloud security, as a more precise diagnosis. Fifty-nine percent of respondents identified critical or significant skills gaps within their teams, up from 44 percent the prior year, and 41 percent identified AI as the single most pressing skill need. Thirty-three percent reported their organization lacked sufficient resources to staff adequately and 29 percent reported being unable to afford personnel with the skills required [33].

However, many organizations do not even get to the point of attempting to fill the roles needed to address the threats they face. The temporal constraint is the organization’s budget cycle. Enterprise security allocates capital annually, through a planning process beginning months before the fiscal year it governs, against a threat landscape that will have changed by the time the money is spent. An investment identified as necessary in one year's third quarter enters planning in the fourth, receives approval if it does in the following year's first quarter, enters procurement in the second, completes vendor selection in the third, and begins deployment in the fourth. Before the first analyst is onboarded, the organization has operated with the gap for over a year. This is not a dysfunctional process but a normal one, reflecting the governance obligations and coordination requirements that large enterprises operate under by design.

There is also a frozen window between the close of one fiscal year and formal approval of the next during which no new spending commitment can be initiated, rarely less than a month and often longer where board ratification is required. During that interval a security team identifying an emerging threat or a newly exploited vulnerability cannot initiate the investment that would address it. The machinery is stopped and the threat is not.

This assumes the best case in staffing enterprise security teams. It assumes the organization is willing to staff that team at levels needed to address the current threat environment. But in practice, organizations often treat requests for additional personnel as a variance from the current baseline. A company could support a small percentage increase in the overall size of the enterprise security team, but requests that would significantly increase the size of the team to meet the new threat environment are treated as unreasonable and dismissed out of hand. As a result the company may be staffed for a threat environment that existed years or even decades ago. In many cases, these budgeting and staffing levels of enterprise security teams were originally set based on assumptions of comparative sufficiency.

The governance properties that make a large enterprise trustworthy, its deliberate capital allocation, its audit trails, its change management discipline, and its accountability structures, are the same properties that make it slower than the economy arrayed against it. The adversary economy has no audit committee, no change advisory board, no procurement regulation, and no fiscal year. What the enterprise built over decades to ensure it is well governed is, from the security function's perspective, also a set of constraints its adversary does not share and has learned to exploit.

Conclusion: The Zebras and the Hyenas

The zebra and hyena framing introduced at the outset has passed through three stages, and each has a specific relationship to the evidence assembled here.

In the first stage, the security posture required of an enterprise was to be a harder target than the weakest of its peers. The adversary population was bounded, its capacity for parallel operation was bounded, and an organization that had invested enough to

distinguish itself from the softest targets would in practice be passed over. Investment was calibrated against peer benchmarks and maturity assessments compared organizations to their industries. The implicit assumption was that adversary attention was a scarce resource that could be diverted.

In the second stage, which the industrialization documented in Section III produced, the threshold rose. The population of capable attackers grew large enough, and the capacity for parallel operation broad enough through affiliate and brokerage structures, that being harder than one's immediate peers was no longer sufficient. The revised formulation was that an enterprise needed to be harder than enough of its peers to satisfy the aggregate appetite of the adversary economy before that appetite reached it. The organizing principle had not changed; the enterprise still needed only to be harder than someone else. But the population of someone else had grown and confidence in sitting above the threshold had eroded.

In the third stage, which the evidence in this paper describes, the framing collapses. When exploit development no longer requires scarce expertise, when convincing deception no longer trades off against volume, when an adversary can learn an organization's own processes well enough to have it authorize the attack itself, and when vulnerability discovery is passing out of the bounds of human research capacity, the assumption that adversary attention is scarce and divertible no longer holds. Capacity sufficient to reach every target of interest, in parallel, on a continuous basis, is the condition the four assumptions were all protecting against, and each of them has now failed. Being harder than one's peers no longer determines whether one is reached, because there is nowhere left to displace the attack to. When capacity is sufficient to reach every target in parallel, the softer target is not visited instead of the harder one. It is visited as well. The only remaining question is whether the defensive posture in place at the moment of contact is adequate against the adversary that arrives. Comparative sufficiency has stopped being a strategy and become a description of how an organization happens to rank while being breached alongside its peers.

Absolute adequacy is a completeness standard, and completeness is precisely what the two compromises negotiated away. Partial completion has been the operating assumption of enterprise defense throughout its history. Multi-factor authentication deployed to most users but not all. Session monitoring built across most authentication contexts but not all. Patching current on most critical assets but not all. Every one of these traded completeness against operational cost on the premise that adversary capacity was bounded and adversary attention was divertible. That premise has expired, and the trade no longer produces the outcome it produced when it was designed. When the adversary economy arrives continuously and in parallel against every reachable surface, the incomplete portion of a defensive posture is not residual risk mitigated by the completed portion. It is the surface through which the intrusion arrives.

For most of the period in which the zebra and hyena logic held, it functioned inside the boardroom as a spending discipline. A board asked to approve a security investment could reasonably ask how the organization's posture compared to its peers, and a favorable comparison offered a defensible reason to hold spending steady. If the organization was not the slowest zebra, the reasoning ran, the marginal investment could wait. This type of comparative sufficiency remains embedded in cyber-risk governance practice today. The most authoritative United States board guidance, the NACD-ISA Director's Handbook on Cyber-Risk Oversight, in its fifth edition published in 2026, still directs boards to ask how significant threats have been to peer organizations, how the organization's control effectiveness compares to industry standards, and how its external vulnerability rating compares to industry benchmarks [34].

The difficulty is not that this guidance is poorly constructed; benchmarking against peers remains genuinely useful for cost discipline and for learning what mature programs look like. The difficulty is that the threat logic which once let a favorable peer comparison double as a spending justification has dissolved. When adversary capacity was bounded and adversary attention was diverted, being harder than one's peers meaningfully reduced the probability of being reached. In the current threat environment it does not. The board that asks what its competitors are doing and takes comfort from a favorable answer is measuring a variable that no longer determines its exposure. The comparison can no longer answer the question that now governs whether the organization is breached, which is not whether its defenses are better than its peers but whether they are adequate, in absolute terms, against an adversary that will arrive regardless of where the organization ranks.

This is not a claim that offense has permanently overtaken defense. The first substantial deployment of frontier vulnerability discovery capability was defensive by design, directed at hardening the software on which the internet depends rather than at attacking it. Thousands of serious vulnerabilities in systemically important code have been found and are being fixed that would otherwise have remained latent, and the defenders who received that capability first genuinely gained an asymmetric advantage. Nor is the situation hopeless. What has happened is that the tools and architectures on which enterprise defense depends were calibrated against assumptions that have expired, that the expiry produces no alarm, and that the compromises which made defense partial by design are no longer survivable.

The response this calls for is not primarily technical. The technical elements, such as continuous session monitoring rather than periodic access review, risk-tiered remediation on the model CISA adopted in Binding Operational Directive 26-04, which scores each vulnerability on exposure, known exploitation, automatability, and impact rather than on severity alone [35], and continuous adversarial testing of controls rather than annual assessment, are all downstream of a governance decision. The decision is to stop treating partial completion as a waypoint on the path to full completion and to recognize it as a defined failure mode with a defined threshold, below which the adversary economy will reach the incomplete portion faster than the enterprise can close it.

The zebras cannot outrun each other any longer, and comparative sufficiency was only ever the practice of hoping they would not have to. The hyenas have industrialized, and the tools that once let a well-run organization judge its own safety by looking sideways at its peers were built on assumptions that no longer hold. What remains is the harder standard: to be, in absolute defensive posture, faster than the hyenas themselves.

References

[1] Verizon, “2026 Data Breach Investigations Report,” Verizon Business, 2026. [Online]. Available: https://www.verizon.com/business/resources/reports/dbir/

[2] The bear is in fact the most common telling in the security literature, where it has already been extended to make this paper’s point: Ben Tomhave characterized the modern threat not as a single bear but as “a drone army of bears,” such that being incrementally faster than the next target no longer suffices. D. Bradford, “Outrunning the Bear,” Advisen, May 8, 2014. [Online]. Available: https://www.advisenltd.com/2014/05/08/36809/

[3] Push Security, “What the Verizon DBIR Tells Us About Breaches in 2026,” June 30, 2026. [Online]. Available: https://pushsecurity.com/blog/verizon-dbir-2026-review

[4] Axonius, “Verizon DBIR 2026: Back to the Fundamentals, Beyond CVEs,” May 20, 2026. [Online]. Available: https://www.axonius.com/blog/verizon-dbir-2026-fundamentals-beyond-cves

[5] National Crime Agency (UK), “The NCA Announces the Disruption of LockBit with Operation Cronos,” Feb. 20, 2024. [Online]. Available: https://www.nationalcrimeagency.gov.uk/

[6] K. Thomas et al., “Framing dependencies introduced by underground commoditization,” in Proc. 14th Annu. Workshop on the Economics of Information Security (WEIS), Delft, Netherlands, June 2015. [Online]. Available: https://econinfosec.org/archive/weis2015/papers/WEIS_2015_thomas.pdf

[7] J. Lusthaus, Industry of Anonymity: Inside the Business of Cybercrime. Cambridge, MA: Harvard University Press, 2018.

[8] L. Ablon, M. C. Libicki, and A. A. Golay, Markets for Cybercrime Tools and Stolen Data: Hackers’ Bazaar, RAND Corporation, RR-610-JNI, 2014. [Online]. Available: https://www.rand.org/pubs/research_reports/RR610.html

[9] Check Point / Cyberint, “Initial Access Brokers Report 2024,” Cyberint Threat Intelligence, Aug. 2024.

[10] KELA Cyber Threat Intelligence, “Infostealer Epidemic Report,” KELA, 2025.

[11] Flare, “2026 State of Enterprise Infostealer Exposure Report,” Flare Systems, 2026.

[12] Mandiant / Google Cloud, “UNC5537 Targets Snowflake Customer Instances for Data Theft and Extortion,” Mandiant Threat Intelligence, June 2024.

[13] Chainalysis, “Crypto Crime Report 2025: Ransomware Activity Down as Victims Refuse to Pay,” Feb. 2025. [Online]. Available: https://www.chainalysis.com/blog/crypto-crime-ransomware-victim-extortion-2025/

[14] Recorded Future, “Initial Access Brokers: The Market That Fuels Ransomware,” Recorded Future Threat Intelligence, 2024.

[15] ESET, “H2 2024 Threat Report,” ESET Research, Dec. 2024.

[16] Palo Alto Networks Unit 42, “The Dual-Use Dilemma of AI: Malicious LLMs,” Unit 42 Threat Research, Nov. 2025.

[17] U.S. Department of the Treasury, Office of Foreign Assets Control, “Designation of Huione Group as a Money Laundering Concern,” May 2025.

[18] Anthropic, “Measuring LLMs’ Impact on N-day Exploits,” Frontier Red Team, June 8, 2026. [Online]. Available: https://www.anthropic.com/research/n-days

[19] Anti-Phishing Working Group, “Phishing Activity Trends Reports, Q1–Q4 2024,” APWG, 2025. [Online]. Available: https://apwg.org/trendreports/

[20] IBM X-Force, “X-Force Threat Intelligence Index 2025,” IBM Corporation, 2025.

[21] CrowdStrike, “2026 Global Threat Report,” CrowdStrike Holdings, Inc., 2026. [Online]. Available: https://www.crowdstrike.com/global-threat-report/

[22] CNN, “Finance worker pays out $25 million after video call with deepfake ‘chief financial officer,’” Feb. 4, 2024. [Online]. Available: https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/

[23] Fortune, “A deepfake ‘CFO’ tricked the British design firm behind the Sydney Opera House in $25 million fraud,” May 17, 2024. [Online]. Available: https://fortune.com/europe/2024/05/17/arup-deepfake-fraud-scam-victim-hong-kong-25-million-cfo/

[24] Microsoft, “Microsoft Digital Defense Report 2025,” Microsoft Corporation, 2025. [Online]. Available: https://www.microsoft.com/security/business/microsoft-digital-defense-report

[25] MGM Resorts International, “Form 8-K Current Report Disclosing Cybersecurity Incident,” U.S. Securities and Exchange Commission, Sep. 13, 2023.

[26] Anthropic, “Project Glasswing: An Initial Update,” May 22, 2026. [Online]. Available: https://www.anthropic.com/research/glasswing-initial-update

[27] Anthropic, “Investigating Three Real-World Incidents in Our Cybersecurity Evaluations,” Frontier Red Team, July 30, 2026. [Online]. Available: https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals

[28] Ponemon Institute and Rezilion, “The State of Vulnerability Management in DevSecOps,” Ponemon Institute, 2022.

[29] Help Net Security, “Companies Keep Getting Breached by Vulnerabilities They Already Knew About,” July 16, 2026. [Online]. Available: https://www.helpnetsecurity.com/2026/07/16/ciso-vulnerability-remediation-gap/

[30] Cybersecurity Dive, “White House Wants to Hold the Software Sector Accountable for Security,” May 10, 2024. [Online]. Available: https://www.cybersecuritydive.com/news/white-house-software-accountable-security/715797/

[31] Cybersecurity and Infrastructure Security Agency and Federal Bureau of Investigation, “Secure by Design Alert: Eliminating SQL Injection Vulnerabilities in Software,” CISA, Mar. 25, 2024. [Online]. Available: https://www.cisa.gov/resources-tools/resources/secure-design-alert-eliminating-sql-injection-vulnerabilities-software

[32] ISC2, “2024 Cybersecurity Workforce Study,” ISC2, 2024.

[33] ISC2, “2025 Cybersecurity Workforce Study,” ISC2, Dec. 2025. [Online]. Available: https://www.isc2.org/Insights/2025/12/2025-ISC2-Cybersecurity-Workforce-Study

[34] National Association of Corporate Directors and Internet Security Alliance, Director’s Handbook on Cyber-Risk Oversight, 5th ed. Arlington, VA: NACD, 2026. [Online]. Available: https://www.nacdonline.org/all-governance/governance-resources/governance-research/director-handbooks/2026-cyber-risk-oversight/

[35] Cybersecurity and Infrastructure Security Agency, “Binding Operational Directive 26-04: Prioritizing Security Updates Based on Risk,” June 10, 2026. [Online]. Available: https://www.cisa.gov/news-events/directives/bod-26-04-prioritizing-security-updates-based-risk