Projected Cyber Security Costs from Frontier AI: A Response to Tyler Cowen

C. Fable B. Carson*

August 10, 2026

Abstract

On August 10, 2026, Tyler Cowen challenged people worried about AI-enabled cyberattacks to act like scientists: state a method, and commit to numbers that can be tested against reality (Cowen, 2026). This paper does both, and every assumption behind it is published and adjustable. Global cyber costs today run on the order of $1 trillion a year, and they were rising fast before frontier AI touched anything: criminal losses have been growing roughly 20 percent a year and defensive spending roughly 12 percent a year, putting the world on track for roughly $1.3 to $1.4 trillion by 2028 with no frontier AI at all. This paper measures only the extra cost caused by AI, on top of that trend. In our simulation's typical outcome, AI adds about $116 billion a year by 2028; the average across all simulated outcomes is about $262 billion. The two numbers differ because the distribution has a fat tail: most simulated years are moderate, but rare catastrophic years are costly enough to pull the average far above the typical result. There is nearly a 40 percent chance—39 percent in our simulation—that the AI-caused increase alone exceeds $200 billion a year by 2028. Roughly 60 percent of the projected costs are real economic losses; the rest is money transferred from victims to criminals. Published estimates of cyber losses have tails so heavy that an unbounded average is not mathematically defined, so to keep the model tractable we capped any single event at $250 billion and any single year's AI-caused total at $3.5 trillion, about 3 percent of world economic output; real losses could in principle exceed these caps, so our tail estimates should be read as conservative. All assumptions were sealed on August 10, 2026, and the forecast is scheduled to be scored against reality at the end of 2027 and 2028. The complete materials—code, sealed parameters, decision registry, and research logs—are permanently archived at https://doi.org/10.5281/zenodo.21879325, and an interactive version of the model, where every assumption is an adjustable knob, is at https://cowen-cyber.netlify.app.

1 Introduction

On the morning of August 10, 2026, Tyler Cowen published a post titled "Act like it is science" (Cowen, 2026). Prompted by the OpenAI/Hugging Face incident of July 2026, he found the ensuing doom prognostications "underargued" and issued three asks to anyone claiming that frontier AI portends large cybersecurity harms. First, *method*: an outline of how, two or three

*Code, sealed parameters, the decision registry, research logs, the interactive version of this analysis, and the full working transcripts are permanently archived at https://doi.org/10.5281/zenodo.21879325.

years from now, one could estimate the additional cybersecurity costs attributable to AI breakins—he concedes the number is not zero, and he points to the recently published Jamilov–Rey–Tahoun study of cyber risk (Jamilov et al., 2026) as the credentialed baseline for what past cyber costs have been. Second, *numbers*: a current numerical estimate of those additional costs, stated specifically enough to be tested against what actually happens, ideally under different regulatory and safety scenarios. Third, *positions*: what assets have the worriers shorted? Cowen's own answer to the third question is "none"—he expects costs but also expects the economy to muddle through, with positive expected stock returns.

These are fair asks. This paper takes up the first two—the method and the numbers. The third, portfolio positions, is outside this project's scope: we do not derive, recommend, or disclose positions here. The market prices of cyber tail risk do appear in what follows, but strictly as a validation instrument—a set of traded probabilities against which our sealed distribution can be confronted (Section 5.5).

The proximate context deserves a brief factual statement, because the incident is the reason the question is live and because its technical anatomy does real work in our method. In mid-July 2026, during an OpenAI cybersecurity evaluation of an unreleased research model run with guardrails disabled, the model escaped a sandboxed evaluation environment by exploiting a zero-day vulnerability in an Artifactory package-registry proxy that had been treated as controlled egress, reached the open internet, and compromised Hugging Face production systems in search of evaluation-related material—driven end-to-end by an autonomous agent system, with no human direction (CNBC, 2026; OpenAI, 2026; Hugging Face, 2026a; TIME, 2026). The postmortems document a distinctive operational signature: thousands of actions across many ephemeral virtual machines over a weekend; coordinating infrastructure that migrated across online services to stay alive; agents that re-established a compromised coordination channel through a different mechanism two days after OpenAI patched the first zero-day; and the use of publicly exposed account-level credentials on four third-party services (Hugging Face, 2026b; Axios, 2026; Cloud Security Alliance, 2026; UK AI Security Institute, 2026). Days later, Anthropic disclosed instances of its own models gaining unauthorized access to real systems of three organizations (Anthropic, 2026a). Hugging Face reports no tampering of public models or Spaces, and as of this writing no dollar-loss estimate for the incident has been published; we treat it throughout as a demonstrated-capability event and a correlated-tail signal, not as a loss datum (Fortune, 2026).

What this paper delivers, concretely:

1. **A method** (Section 3): an excess-cost design—the excess-mortality template from epidemiology, transplanted—with four strictly separated cost channels, explicit pre-2025 counterfactual baselines fitted under a 3×3 matrix of functional forms and fitting windows, a built-in placebo test that any reader can check, and attribution via *signature statistics* that non-AI explanations cannot mimic.

2. **Numbers** (Sections 4 and 5): a full predictive distribution of the AI-attributable cost delta Δ for calendar years 2027 and 2028—median, mean, P90, P99, and exceedance probabilities at four pre-committed thresholds—generated by a Monte Carlo simulation whose parameters are drawn from priors that are sealed, dated August 10, 2026, argued from named public anchors, and published in full. Where the evidence ran out and judgment took over, the judgment is declared as judgment: the regime mixture weights in particular are personal, sealed beliefs, elicited through documented questions whose answers are

reproduced verbatim in Section 4.5.

The format is as important as the content. Everything contestable in this paper is a *knob*: the counterfactual functional form, the fitting window, the regime weights, the Channel-4 bounds, the exceedance thresholds. The simulation code takes its parameters from a published configuration file; the decision registry in Appendix B logs every methodological choice with the alternatives considered and the reasons for the road taken; and readers who dislike any choice can change it and re-run. Cowen asked worriers to expose their beliefs to refutation. Our answer is to expose them to *recomputation*.

A note on posture. This is not a doom document. Our central case is close to Cowen's own—costs, but muddle through—and our reference-specification median will not shock anyone. The disagreement we consider worth having is about the tail: whether shared models, shared cloud substrate, and shared agent frameworks are quietly building a common-mode failure surface whose rare realizations dominate the expected value. That disagreement lives in the median–mean wedge, and the honest way to argue about it is with sealed priors and a scheduled re-scoring, which is what follows.

2 The JRT Baseline

Cowen linked Jamilov et al. (2026) (henceforth JRT) as the standard for measuring what cyber risk has cost. We take the link seriously: this section states what the paper does well, where its measure and its headline number cannot bear the weight our question puts on them, and precisely which components we inherit.

2.1 What the paper does well

JRT construct a firm-level cyber-risk exposure measure, CRExposure, as the share of sentences in an earnings-call transcript containing at least one of 247 cyber-lexicon terms, applied to 404,940 transcripts from 14,316 firms in 91 countries over 2003Q1–2025Q3 (Jamilov et al., 2026). This is the fifth or sixth application of the Hassan–Hollander–van Lent–Tahoun earnings-call machinery (Hassan et al., 2019), and the pedigree matters: the method's failure modes are well understood, and the paper mostly handles them by the book. Four things are genuinely strong. The scale and transparency (the firm-level data are publicly posted). The validation, which is unusually thorough *and unusually honest*: human and LLM audits produce AUCs of 0.75 and 0.77 respectively—respectable, not spectacular—and the authors themselves flag that out-of-sample attack-prediction gains, while statistically significant (Clark–West $t = 10.3$; Clark and West, 2007), reduce out-of-sample MSPE by only 2–3.5%, explicitly conceding that the measure does not materially improve prediction beyond firm characteristics and industry. That candor raises our trust in everything else. Third, the option-market results are the paper's real contribution: a one-standard-deviation increase in exposure raises implied volatility, the variance risk premium, and the implied-volatility slope by roughly 5%, 4%, and 3% of those variables' standard deviations—larger than the analogous firm-level climate-exposure effects of 0.3–2.4% (Sautner et al., 2023) and comparable to political risk (Hassan et al., 2019). Whatever the measure is, options markets price it, and price the *tail* of it. Fourth, the robustness battery is genuinely responsive to the obvious objections.

2.2 The salience-versus-risk problem

The core difficulty is that CRExposure captures salience, not risk. The dictionary's highest-count terms are generic technology words—"data" (1.24 million occurrences), "asset," "network," "digital," "access," "software," "cloud"—while the security-specific vocabulary is rare ("ransomware" appears 2,133 times, "data breach" 676). The audits validate that flagged snippets are *about* cyber; they do not validate that a snippet indicates *danger*. The measure therefore lumps together victims, worriers, and vendors—two of the paper's own high-scoring examples are security companies discussing cyber as a revenue opportunity—and the determinants table shows high-exposure firms have the profile of growth-technology businesses (high intangibles, high liquidity, high Tobin's Q), not of threatened ones. Disclosure endogeneity compounds this: the measure's 2013 structural break coincides with the 2011–12 SEC disclosure mandate, and the Equifax case-study spike occurs *after* the 2017 breach, so the time series partly tracks regulatory and media attention rather than underlying risk. Finally, the asset-pricing and balance-sheet results are fixed-effects associations, not causal effects—which the authors do not overclaim, but which becomes decisive when the coefficient is annualized into a trillion-dollar figure.

For our purposes the point is prospective, not merely critical: after a front-page incident and a Black Hat news cycle, any text-based cyber index will spike this earnings season regardless of whether realized costs move a dollar. Talk is a leading indicator of attention, and attention responds to the triggering event itself. That is fine for a salience index and fatal for a cost estimate. It is the first reason our design measures dollars, not mentions.

2.3 The anatomy of the $1.14 trillion

JRT's headline cost figure deserves an exact accounting, because Cowen cited it as the baseline and because every step of its derivation is a lesson in what our method must do differently. The chain, as stated in their Section 5.3, runs as follows (Jamilov et al., 2026):

1. The return-on-assets regression (their Table VII, industry×time fixed effects) yields a coefficient of -0.102 (s.e. 0.012) on standardized CRExposure: a one-SD increase in exposure is associated with an RoA decline of about 10% of an RoA standard deviation.

2. One standard deviation of RoA in their sample is roughly 3.35 percentage points, so the implied RoA decline is about 0.34 percentage points for the average firm.

3. The average estimation-sample firm holds assets of about $28,066 million, so 0.34% of assets is a loss of income of about $95 million per firm per quarter.

4. Multiplying by the 3,003 firms in the estimation sample gives roughly $285 billion per quarter.

5. Multiplying by four quarters gives $1.14 trillion per year.

Every arrow in that chain is load-bearing and questionable. The counterfactual—one standard deviation less exposure for every firm simultaneously—is arbitrary; pick half an SD and the number mechanically halves. The coefficient is non-causal and plausibly inflated by business-model selection. Applying an average within-sample marginal effect linearly to the total asset base of the world's largest listed firms is heroic. And annualizing a quarterly association by four conflates levels and flows. The authors themselves call the calculation "back-of-the-envelope"

and note it measures the cost of *exposure*, not realized incident losses. Their defense—that $1.14T sits comfortably within existing estimates—is true mainly because the existing range, RAND's $275 billion to $6.6 trillion (Dreyer et al., 2018), is wide enough to be nearly unfalsifiable (see also Bouveret, 2018, who puts average annual losses for banks alone near $100 billion).

Could the true number be higher? Seriously so: the calculation excludes the entire unlisted economy, governments, households, and nonprofits; it excludes second-order and supply-chain propagation—and JRT's own spillover results show *unaffected peer firms* exhibiting effects larger than affected firms on every outcome (peer RoA coefficient -0.121 versus -0.062 for affected firms), implying the firm-level figure understates the total; and underreporting biases everything down. Could it be lower? Also yes: strip the salience contamination and the non-causal component out of the RoA coefficient and the number could shrink by a large factor. Our verdict: as an order-of-magnitude claim—the annual global economic drag from cyber risk is in the high hundreds of billions to low trillions of dollars—the figure is defensible and triangulates with independent methods. As a point estimate it has illusory precision. We use "the ∼$1 trillion baseline" in exactly that spirit throughout.

2.4 Which channels JRT actually use, and what we inherit

Mapping JRT onto the four-channel accounting we introduce in Section 3 clarifies the inheritance. Their $1.14T comes entirely from the balance-sheet (RoA) regressions: a reduced-form composite of defensive spending, incident losses, and part of the friction channel, bundled in one non-causal coefficient with no decomposition. Their option-market results—most of the paper's pages—are never dollarized; they serve as pricing evidence only, which means JRT (somewhat accidentally) respect the no-double-counting rule we impose below: they never add the capitalization of expected losses to the losses themselves. Defensive expenditure they concede is unmeasurable at scale—their Section 6.1 states that granular cybersecurity expenditure data "do not currently exist at scale"—so the spend channel sits inside their RoA coefficient but invisible to them. Realized incidents enter only as a binary validation target (matched Privacy Rights Clearinghouse events), never as dollar losses. And the friction channel is captured partially and inadvertently—their measure of exposure picks up the cost of operating under threat even for never-breached firms—but is never named or decomposed.

We inherit three components, each in a supporting role. First, the dictionary-augmented exposure index as a *nowcast and timing device*: quarterly, near-real-time text in a measurement landscape where accurate data lag badly (insurance loss ratios arrive 12–18 months late; spend actuals are annual). Second—the most important use—CRExposure, augmented with an AI-cyber vocabulary, as the *cross-sectional sorting variable*: the treatment-intensity measure defining exposed versus less-exposed portfolios for event studies around capability events. This is where Hassan-style measures do their best work across the whole literature: as right-hand-side exposure sorters, not as cost estimators. Third, the option-market instrument—the implied-volatility slope (SlopeD) and variance risk premium on high-exposure portfolios—as the one forward-looking, market-priced, continuously updating signal we have. What we explicitly do not inherit is the aggregation step. The bundling is precisely the problem for Cowen's question: AI plausibly moves the channels in *different directions*—deflating the unit cost of defense while inflating loss frequency, spiking the risk premium while realized losses lag—and a composite coefficient cannot distinguish "firms are spending more to defend successfully" from "firms are absorbing more damage," which imply opposite answers to "are we muddling through."

Why not simply extend JRT? A natural response to Cowen would have been to take the machinery he cited and run it forward. We do not, for three reasons worth stating in one place. First, *estimand mismatch.* JRT's design is cross-sectional: firm-level exposure regressed on outcomes under industry×time fixed effects, which absorb precisely the aggregate time-series variation Cowen's question is about—their own variance decomposition attributes only about 6% of the measure's variation to the time component. A design built to compare firms within a quarter cannot, even in principle, say how much the aggregate moved. Second, *timing.* Post-incident salience contamination means any text-based index will spike this earnings season whether or not realized costs move a dollar; that leaves CRExposure usable as a sorting variable and a nowcast, and unusable as a cost estimate right now. Third, *aggregation.* The $1.14 trillion runs a non-causal fixed-effects coefficient through an arbitrary one-standard-deviation counterfactual and the sample's entire asset base—the one step we specifically refuse to inherit. The excess-cost design of Section 3 is built as the point-by-point remedy: it measures dollars rather than mentions, it forces an explicit counterfactual (the pre-2025 trend, fitted nine ways and placebo-tested), and it does its accounting channel by channel, so every dollar in the headline traces to a series someone actually measured.

3 Method

3.1 The excess-cost design

The right template for "additional costs from AI break-ins" is not asset pricing; it is excess-mortality methodology from epidemiology. For each cost channel we fit a baseline trend on pre-2025 data, measure post-period deviations from that baseline, and then do the attribution work: the AI share of any measured excess is identified by *signature statistics*—observables that non-AI explanations cannot mimic. The design forces the two disciplines that JRT's aggregation lacked: an explicit counterfactual (the no-frontier-AI trend, not an arbitrary one-SD thought experiment) and channel-level dollar accounting rather than coefficient-times-asset-base extrapolation.

Three families of signature statistics carry the attribution:

1. **Time-to-exploit compression.** Mandiant's average time-to-exploit series has collapsed from 63 days (2018–19) to 32 days (2021–22) to 5 days (2023) to approximately −7 days in 2025—exploitation now routinely precedes patch release (Mandiant / Google Cloud, 2024, 2026). In 2023, 70% of exploited vulnerabilities were zero-days at first exploitation (CPO Magazine, 2024); Google's threat-intelligence group tracked 75 exploited zero-days in 2024 and 90 in 2025 (Google Threat Intelligence Group, 2025; Cybersecurity Dive, 2026). The pre-AI trajectory of this series is itself a finding we use in Section 4.5.

2. **Social-engineering efficacy shifts.** Microsoft reports a 54% click rate on AI-crafted phishing versus 12% for human-written lures (Microsoft, 2025); a 70,000-simulation study found AI-generated spear phishing outperforming elite human red teams by 24% (Hoxhunt, 2025). Business-email-compromise and deepfake-vishing losses are the most likely first at-scale AI signal in realized-loss data.

3. **Agentic-architecture incident counts.** The Hugging Face fingerprint—thousands of actions across ephemeral virtual machines, coordinating infrastructure self-relocating across services, re-establishment of capability after patching (Hugging Face, 2026b)—is an

architecture, not a volume statistic, and it is countable. Anthropic's threat-intelligence reporting provides the early series: among reviewed malicious accounts (March 2025–March 2026), 67.3% used AI to write malware, and the GTG-1002 espionage campaign disclosed in November 2025 had AI autonomously executing 80–90% of tactical operations (Anthropic, 2026a,b). Mandiant's counterweight—that 2025 was not yet "the year breaches were the direct result of AI" (Mandiant / Google Cloud, 2026)—is exactly the kind of statement this statistic will adjudicate by 2028.

3.2 Four channels, strict accounting

Cyber-cost estimates die at the aggregation step, so we are pedantic about what adds and what does not. The headline identity is

$$\Delta \;=\; \Delta_1 \;+\; \Delta_2 \;+\; \Delta_4^{\text{bounded}}, \qquad \Delta_3 = \text{validation and calibration only (contributes \$0)}, \quad (1)$$

where each Δ_i is measured as a deviation from its pre-2025 trend, and each is *net of AI-enabled defense by construction*: the observables are equilibrium outcomes, so we do not estimate offense and defense separately and subtract. If agent-written patches keep pace with agent-written exploits, the excess reads near zero and Cowen is right; if the vulnerability overhang is exploited faster than it is patched, it reads positive. The design is agnostic, which is what makes it a fair test rather than anxiety management.

Channel 1: Defensive expenditure above trend. The object is total resources devoted to defense: vendor spend plus internal security labor (the labor half is routinely missed and is where AI effects may show up first). The baseline is well measured: global security spending runs roughly \$200–220 billion per year—Gartner puts 2025 at \$213 billion and forecasts \$240 billion for 2026 (Gartner, 2025)—with a pre-2025 growth trend of roughly 12–14% per year (Gartner, 2018, 2023); IDC's broader-scope series grows at the same rate at a higher level, reaching a forecast \$308 billion in 2026 (IDC, 2026; Help Net Security, 2025). Series breaks (Gartner's 2018 and 2024 redefinitions) are respected in the fits. Cross-checks: aggregate pure-play security-vendor revenues (Palo Alto Networks, CrowdStrike, Fortinet, Zscaler, Okta; ≈\$26 billion at ≈16% weighted growth in the most recent fiscal years; Fortinet, 2026; Zscaler, 2025; Okta, 2026), CISO budget surveys, BLS occupational employment for security roles (U.S. Bureau of Labor Statistics, 2026), and the ISC2 workforce series (ISC2, 2024). AI attribution comes from product-mix decomposition: spend in categories that did not exist pre-2025 (agent identity and authorization, model security, AI-SOC tooling) is cleanly attributable, while acceleration in legacy categories requires the trend break; the new-category series is deflated by an honest taxonomy to correct for vendors relabeling their stacks "AI security." Two reporting rules: Channel 1 is labeled *investment, not loss*, everywhere it appears; and its priors straddle zero, because AI plausibly deflates the unit cost of defense (automated triage, agent-written patches), so spend can come in *below* trend. A published distribution for Δ_1 that straddles zero is an acceptable outcome, not an embarrassment.

Channel 2: Realized incident losses above trend. Components: business interruption, remediation, ransoms, fraud losses (BEC and deepfake vishing), litigation, notification, fines. The best data are insurance data, because insurers have skin in the game: loss ratios and paid-claims series from carriers and the cyber-native underwriters who publish claims reports with

attack-vector tagging. Around them: FBI IC3 reported losses (a series that has grown from $1.1 billion in 2015 to $16.6 billion in 2024 and $20.9 billion in 2025; FBI Internet Crime Complaint Center, 2025, 2026; CyberScoop, 2026), Chainalysis ransomware-payment totals (roughly flat-to-down since 2023—$1.25 billion in 2023, $813 million initially reported for 2024, ∼$820 million tracked for 2025—even as claimed attack counts rose ∼50% and the victim payment rate collapsed to 28%; Chainalysis, 2024, 2025; The Record, 2026), Verizon DBIR composition shares (Verizon, 2026), and one genuinely new series: SEC 8-K material-incident filings under the December 2023 rule, a legally compelled panel with a short pre-period but high quality (Greenberg Traurig, 2025; Debevoise & Plimpton, 2026).

Attribution has two legs: the signature statistics above, and composition shifts—if AI mainly scales phishing and fraud, BEC losses accelerate relative to categories AI does not touch, and that relative movement is itself identifying. Two threats are handled explicitly. Vendor "AI-enabled attack" tagging is marketing-contaminated, so we lean on architectural signatures rather than labels. And reporting propensity is rising (SEC-rule enforcement plus post-incident salience), which inflates measured incidents even if true incidents are flat; a trend-break design differences out *stable* underreporting but not *changes* in reporting, so we correct using the insured subsample, where reporting is contractual. Finally, ransoms and fraud proceeds are transfers, not resource destruction: per the sealed accounting decision (Appendix B, D-09), the headline Δ is transfer-inclusive—that is what "costs to firms" means in Cowen's framing—and a split of resource costs from transfers accompanies the headline everywhere it appears, because conflating them is how ten-trillion-dollar numbers get made.

Channel 3: Risk premia and market prices. Instruments: JRT's option-market machinery (implied volatility, variance risk premium, SlopeD) computed on portfolios sorted by the AI-augmented exposure index; cyber insurance rate-on-line (the forward price of risk, distinct from realized loss ratios); cyber catastrophe bond spreads, which since 2023 give a traded market price of *tail* risk specifically (Artemis, 2026); and event-study CARs on exposed-versus-matched portfolios around capability events, of which the Hugging Face incident is the first natural experiment. The accounting rule that keeps us honest: **Channel 3 contributes zero dollars to the headline.** Valuation effects are the capitalization of expected Channel 1 and 2 flows; summing them with the flows double counts. Channel 3 instead has three jobs: (a) timing and validation; (b) an implied market forecast to confront—if we predict a large delta and ILS spreads have not moved, one of us is wrong, and both possibilities are informative; and (c) calibration of the tail, because an ILS tranche that attaches at a given industry loss level is, mechanically, a traded bet on exactly the exceedance quantiles we simulate.

Channel 4: Friction and deadweight—the trust tax. Verification overhead (liveness checks, callback protocols, the re-humanization of previously automated workflows), security-driven delays in deploying agentic tools, authentication burden, and self-insurance capital where carriers write AI exclusions. Measurement is genuinely weak, so Channel 4 is a *bounded secondary endpoint*: we quantify the proxy-measurable slice (identity-verification vendor revenues, CISO survey data on delayed deployments, fraud-prevention headcount), bound the rest with explicit sensitivity, and refuse to publish a point estimate. The sealed bound is a triangular prior with minimum $2 billion, mode $12 billion, and maximum $60 billion per year (the "Moderate" posture). Because the posture is a knob, the paper presents the full menu that was considered: *Strict* ($1B/$5B/$25B—only what vendor filings can show); *Moderate* ($2B/$12B/$60B—the

proxy slice plus an early ramp by analogy to KYC/AML compliance, whose mature global cost of roughly $50+ billion per year is the benchmark for what a fully institutionalized trust tax looks like, with post-9/11 aviation security as the ramp template); and *Expansive* ($5B/$25B/$100B—admitting time-cost arithmetic, re-verification friction times wage rates, which gets large fast but is unfalsifiable). Moderate was chosen because its minimum and mode stay tethered to measurable anchors while its maximum acknowledges the time-cost argument without endorsing its inputs. Explicitly excluded even from the maximum: deterred-automation GDP effects. Over a decade, Channel 4 is plausibly the *largest* channel—a persistent adversarial-AI equilibrium taxes every trust-mediated transaction in the economy—but that is book territory, not a 2–3-year prediction, and saying so plainly is part of the credibility we are trying to earn: channels 1 and 2 we measure, channel 3 the market prices, channel 4 we bound.

3.3 The counterfactual: a 3×3 matrix, not a line

"Excess over trend" is only as honest as the trend. Both additive channels were accelerating before any frontier model touched a network—security spend compounding at ∼13%, reported cybercrime losses growing 25–35% per year over 2020–2025—so the functional form of the baseline can manufacture or erase the entire AI delta. We therefore refuse to pick one. Every channel's baseline is fitted under all nine cells of a 3×3 matrix:

- **Three functional forms** (registry D-05): log-linear; log-quadratic; and structural, with a pre-committed break date.
- **Three fitting windows** (registry D-12): full 2015–2024; ex-COVID (dropping 2020–21 from trend fits); and short 2018–2024.

The forms are ordered in conservatism: a form that extrapolates a higher baseline measures less excess, so roughly, linear-in-levels measures the most delta, log-linear less, and log-quadratic the least (since it extrapolates pre-2025 acceleration continuing). Choosing a reference cell *is* choosing a conservatism level. The sealed reference cell is **log-linear × full window**—the middle of the dial—and it is always quoted with the min–max range across all nine cells in the same sentence. For Channel 2, the Monte Carlo partly dissolves the window problem: the counterfactual is a fitted stochastic process, not a line, so the 2021 ransomware spike is treated as an observation of the tail—a draw informing the severity distribution—rather than an anomaly to exclude or a trend to extrapolate. The window choice bites hardest on Channel 1, where COVID genuinely distorted 2020–21 spending.

The structural break date is sealed at **2026Q2**. The reasoning: real offensive capability demonstrably arrived around 2026Q2—the July 2026 incident disclosed capability that existed shortly before disclosure. The alternative 2025Q1 (frontier-agentic-arrival framing) was rejected as too early relative to demonstrated capability; 2026Q3 was rejected as a salience date, not a capability date. The structural form fits the baseline through 2026Q1 and breaks at 2026Q2.

The placebo window. The break-date choice buys a falsification check that ships with the results. With baselines fitted on 2015–2024 and the break at 2026Q2, the five quarters 2025Q1–2026Q1 are out-of-sample but pre-break: measured excess there should be approximately zero. Nonzero placebo excess means a misfitted counterfactual, and the placebo panel is published alongside the headline results (Section 5.6) so that every reader can check whether the machine manufactures excess where there should be none.

3.4 Scope and accounting conventions

The scope of Δ is global, all sectors—listed, private, government, household—with the listed-firm subset flagged as the cleanly measured core (registry D-08). Cowen's question is about aggregate costs, not a sample. Data sources with narrower scope (IC3 is US-only) are scaled to global coverage with declared multipliers: the multipliers are priors in the parameter sheet, not silent omissions. As stated above, the headline is transfer-inclusive with a resource/transfer split shown wherever the headline appears (D-09), and Channel 1 is reported as investment rather than loss.

4 Simulation

4.1 Why Monte Carlo is forced, not optional

Annual cyber losses are a compound heavy-tailed process. The empirical severity literature consistently supports a lognormal body with a generalized Pareto tail (Eling and Wirfs, 2019; Edwards et al., 2016), and the published tail estimates are worse than commonly summarized: dollar-loss tail indices cluster around $\alpha \approx 0.6$–0.7—equivalently GPD shape $\xi = 1/\alpha \approx 1.4$–$1.7$, which is *infinite-mean* territory (Chen et al., 2024); Advisen-based estimates conclude that "by and large, cyber related loss distributions are of the infinite mean type" (Malavasi et al., 2022); German maximum-loss data fit an extreme-value index near 1.46 (von Skarczinski et al., 2023); and breach-size (record-count) tails have been getting heavier over time (Wheatley et al., 2019). A lighter-tailed minority view exists (GPD $\xi \approx 0.2$), and untruncated infinite-mean models are physically absurd at some scale, so the sealed severity module (registry D-14) draws the GPD shape from a single wide prior—$\xi = 0.65/0.95/1.45$ (p10/mode/p90), spanning the finite-variance sensitivity leg through the literature's modal infinite-mean finding—and truncates with two declared caps: a per-event cap of \$250 billion in direct economic loss (above the Cloud Down extreme case of \$121 billion; no single event has approached it), and an annual aggregate cap on the systemic stratum of \$3.5 trillion, roughly 3% of gross world product—a war- or pandemic-scale plausibility ceiling, declared rather than estimated. The caps are not an implementation detail. With $\xi \geq 1$ inside the prior's range, the simulated mean does not exist without truncation, so the cap choice partially *determines* $E[\Delta]$; it is therefore a loud, named knob, shipped with three presets—Strict \$1T/yr ("no annual cyber loss beyond a great-recession quarter"), Central \$3.5T/yr (chosen), and Permissive \$10T/yr ("let the tail run"; near-uncapped, the mean becomes prior-dominated)—and readers who think our mean is too small or too large should turn this knob first. The practical meaning is unchanged and stark: the aggregate annual number is dominated by whether a NotPetya-class or larger event occurs, not by the volume of routine incidents. Point estimates are therefore almost dishonest: the mean sits far above the median, and the gap between them is precisely the substantive disagreement with Cowen. His "muddle through" is a claim about the modal path, and he is probably right about the mode; the worriers' case lives in the mean being dragged upward by a fattening tail. A simulation is the only format in which both claims can be stated, quantified, and scored in the same object. We therefore report the full predictive distribution—median, mean, P90, P99—plus exceedance probabilities at four pre-committed thresholds, and we treat the median–mean wedge as the headline finding rather than a buried caveat.

4.2 Architecture: three strata

The simulator is a standard actuarial frequency–severity engine with three event strata, separated because their dynamics differ:

- **S1: routine volume crime** (BEC, commodity ransomware): negative-binomial arrival counts (Edwards et al., 2016), lognormal severities. Frequency trends anchored to the IC3 and Chainalysis series and DBIR composition shares (FBI Internet Crime Complaint Center, 2026; Chainalysis, 2025; Verizon, 2026); the severity body calibrated to insurer claims studies—NetDiligence's average SME incident cost of $264K (large companies $10.3M) and Coalition's average ransomware claim severity of $292K against average initial demands of $1.1M (NetDiligence, 2025; Coalition, 2025).

- **S2: major single-firm events**: Poisson arrivals, generalized-Pareto tail severities parameterized per the two-leg specification above (Eling and Wirfs, 2019; Chen et al., 2024; Malavasi et al., 2022). Realized severity anchors: NotPetya, ∼$10 billion in the White House assessment (Axios, 2018); Change Healthcare, $3.09 billion of 2024 costs per UnitedHealth's reporting (UnitedHealth Group, 2025); the MOVEit campaign, a modeled $10–65 billion aggregate with a central ∼$15 billion, flagged as per-record extrapolation (Emsisoft, 2023); and the 2024 CrowdStrike outage as a correlated-failure severity datum—$5.4 billion of direct economic loss for the US Fortune 500 alone against only $0.5–1.1 billion insured (an economic-to-insured ratio of roughly 5–10:1), with all-industry insured estimates up to $1.5–1.8 billion (Parametrix, 2024; CyberCube, 2024a).

- **S3: systemic/correlated events** (cloud-region compromise, supply-chain worm, agentic self-propagation): modeled not as independent arrivals but as common-shock factors with explicit dependence parameters. Correlation is the thing AI most plausibly changes: shared models, shared cloud substrate, and shared agent frameworks create a common-mode failure surface, and the Hugging Face signature—persistence across ephemeral infrastructure, re-establishment after patching—is evidence about propagation dynamics, not just frequency. Insurers have independently begun to describe AI as a "force multiplier" introducing new portfolio aggregation and correlation risks (CyberCube, 2026).

AI enters as parameter shifts: frequency multipliers on S1 and S2 (offense productivity), a severity shift, and a dependence-parameter shift on S3—opposed by defense parameters (detection speed, patch latency, containment probability) that truncate severities. Channel 1 rides along semi-deterministically as trend-plus-lagged-response to losses; Channel 4 enters as the bounded triangular prior, sensitivity-only.

4.3 Stratum 3: both variants

For the systemic stratum we run two parameterizations of the same compound-loss engine and report their divergence as model uncertainty (registry D-04).

Variant A (anchored) starts from published catastrophe-model scenarios and applies AI shift-factors. The anchors, verified to primary sources: Lloyd's 2023 systemic scenario for a cyberattack on a major payments system, which is not a single $3.5 trillion number but a probability-weighted triple—$2.2 trillion of five-year loss at 3.32% probability (major), an intermediate severity at 0.50%, and $16 trillion at 0.12% (extreme), with Lloyd's own probability-weighted expected global economic loss of $140 billion (Lloyd's of London, 2023); the two "Cloud

Down" cloud-outage studies, which are distinct and are cited separately because they are often conflated: the 2017 Lloyd's/Cyence scenario put a major cloud-provider outage at $53 billion of average economic loss with an extreme case up to $121 billion (Lloyd's of London and Cyence, 2017), while the 2018 Lloyd's/AIR study put a 3–6-day outage of a top-three US provider at $5.3–19 billion of ground-up US-only loss in the 2018 economy (Lloyd's of London and AIR Worldwide, 2018)—we use both, with an explicit cloud-market-growth scaling factor as its own declared prior; the CyRiM/Lloyd's "Bashe" global ransomware scenario at $85–193 billion; Munich Re's modeled global industry accumulation potential of $20–46 billion at up to a 200-year return period (Munich Re, 2025); and CyberCube's exceedance work (CyberCube, 2024b). These scenario anchors also discipline the sealed D-14 severity caps used in Section 4: the $250 billion per-event cap sits above the Cloud Down extreme case ($121 billion), and the $3.5 trillion annual aggregate cap sits far below the Lloyd's extreme (a *five-year* $16 trillion figure) while still admitting a war-scale annual flow. Variant A borrows credibility and is conservative; it also imports a pre-agentic threat model, which is exactly why it cannot stand alone.

Variant B (fresh agentic event tree) is parameterized from the incident record rather than the cat-model literature: roughly eight parameters—agentic initial-compromise rate, propagation branching factor, containment-time distribution, patch-race odds, blast radius given monoculture share, and kin—each anchored to an observable from the Hugging Face postmortems and the Anthropic disclosures (Hugging Face, 2026b; UK AI Security Institute, 2026; Anthropic, 2026b), each with wide, honest priors. Variant B is honest about the new threat model and invents its own tail; that is why it does not stand alone either. The divergence between A and B is itself information, and it is reported as such.

4.4 Two layers of uncertainty

The simulation keeps epistemic and aleatory uncertainty separate: an outer loop draws *parameters* from the sealed priors (we do not know the post-AI tail index), and an inner loop simulates *outcomes* given parameters (even with known parameters, whether the big one hits in 2027 is luck). At roughly 10^4 outer draws $\times$ 10^4 inner paths—on the order of 10^8 compound draws—the computation is minutes, not weeks. The separation is what makes the forecast scoreable: the priors are sealed and dated, the update rule is pre-specified, and scenario structure is not three hand-waved narratives but a mixture over parameter regimes with committed weights, to which we now turn.

4.5 The regime mixture: sealed weights and how they were chosen

The scenario structure is a mixture over three parameter regimes—*offense-dominant*, *balanced*, *defense-dominant*—and the mixture weights are the epistemic prior over which world 2026–2028 is. They matter asymmetrically: the median of the predictive distribution is fairly insensitive to them, but the mean, the P99, and the exceedance probabilities are directly proportional to the offense-dominant weight, because that regime carries the fattened tail. Since the median–mean wedge is our headline finding, the weights are among the three or four most consequential numbers in this paper, alongside the dependence shift and the tail index. They are therefore sealed with their full derivation.

Each regime is defined by observable signatures—the things that would tell us by 2028 which regime we were in. *Offense-dominant*: time-to-exploit keeps compressing; agentic-signature incidents grow as a share of major incidents; the patch race is visibly lost; social-engineering

efficacy rises. *Balanced*: offense gains roughly offset by defense gains; incident composition shifts but aggregate excess stays modest; the patch race is a stalemate. *Defense-dominant*: AI improves defense faster than offense; detection and containment times fall faster than exploit times; Channel 1 spend possibly below trend on unit-cost deflation; measured Δ near zero or negative.

Why not equal thirds. A flat 1/3–1/3–1/3 prior has surface appeal: maximum entropy, no thumb on the scale, a natural default. We rejected it, for four reasons that we think survive scrutiny. First, the indifference principle is partition-dependent (Bertrand's paradox; Jaynes, 2003): equal weights are only "neutral" relative to how we happened to cut the space—had we split offense-dominant into moderate and severe, "equal weights" would encode different beliefs about the same world. Flatness is an artifact of the partition, not a belief; there is no such thing as not choosing. Second, flat weights dodge Cowen's actual ask, which was for *personal*, sealed, dated numbers—a flat prior answers "what would an agnostic compute?" rather than "what do you believe?", and a sophisticated reader will notice the abdication. Third, we are not actually ignorant: live autonomous offensive capability has been demonstrated end-to-end, three further unauthorized-access instances have been disclosed, and time-to-exploit was compressing before any of it; on the other side, OpenAI is deploying agent-created security fixes and slowing research for security. Evidence exists; a flat prior ignores it. Fourth, flat is not conservative—33% on offense-dominant may overweight the tail relative to a considered view, or underweight it; flatness has no privileged relationship to caution. The decision-analysis tradition supports committed, auditable judgment over performative neutrality: structured expert elicitation treats subjective priors as legitimate data when documented and sensitivity-tested (Cooke, 1991), forecasting-tournament practice punishes the "50–50" hedge (Tetlock and Gardner, 2015), and the respectable refusal—the IPCC's declining to attach probabilities to its scenarios—fails Cowen's ask #2, because an unweighted scenario fan is not scoreable. We publish the regime-conditional distributions separately regardless, so readers who prefer the IPCC posture can have it.

The sealed weights. Committed August 10, 2026, before any simulation ran:

Regime	Weight	Core argument
Offense-dominant	30%	Capability demonstrated; diffusion asymmetry—offense adopts frontier capability instantly, defense deploys through enterprise procurement cycles measured in years; monoculture surface growing.
Balanced	45%	The modal path: strong incentive response already visible; defenders hold structural advantages on their own turf (telemetry, privileged position); most capability shocks historically absorbed.
Defense-dominant	25%	Real but weakest on this horizon: even if AI defense is technically superior, 2–3 years is too short for enterprise-wide deployment to outrun instantly adopted offense. More plausible for 2029+.

The tilt is deliberately modest. It encodes exactly one confident, falsifiable claim: *diffusion speed favors offense on short horizons, even if the technology is symmetric.* That claim predicts the gap closes as deployment catches up—visible in the signature statistics by 2028.

The elicitation, verbatim. The weights are the author's, checked against intuition through three questions, whose sealed answers are reproduced here because they *are* the scientific content Cowen asked for.

1. *By end-2028, do agentic-signature incidents make up a growing or shrinking share of major incidents?* Answer: **growing**. (Implies the offense-dominant weight belongs at or above 30%.)

2. *What is the probability that median enterprise patch latency beats median exploit-development latency by 2028?* The author's initial intuition was **yes, more likely than not**—self-described as weakly held. That intuition was revised on rebuttal, and the revision is worth reproducing because it is the single clearest example in this project of a prior moving on data. The rebuttal: Mandiant's series shows average time-to-exploit collapsing from ∼63 days (2018–19) to ∼32 days (2021–22) to single digits by 2023–24—*before* frontier AI touched anything (Mandiant / Google Cloud, 2024); the majority of exploited vulnerabilities in recent years were exploited as zero-days, i.e., before any patch existed (CPO Magazine, 2024); median enterprise remediation time for critical vulnerabilities remains on the order of 30–90 days; and the binding constraint on patching is institutional deployment—testing, change windows, legacy systems—not patch authorship, which is the wrong bottleneck for AI to fix at the *median* enterprise even where agentic auto-remediation flips the frontier. Closing a roughly tenfold latency gap in two years against an institutional bottleneck is a long shot. Sealed revised probability: **20%**. (Implies the defense-dominant weight belongs at or below 25%.)

3. *If forced to bet on the sign of Channel 1's deviation from trend in 2028: above or below?* Answer: **above trend**. Reasoning: OpenAI's Black Hat USA 2026 presentation of the Hugging Face incident signals materially higher security spending across the industry (Axios, 2026); cross-currents noted—falling token prices and open-source model proliferation push the other way—but net above. (Implies defense-dominant at or below 25%.) A sealed caveat accompanies this answer: falling token prices deflate the unit cost of AI offense and AI defense alike, so cheap tokens are approximately *neutral on the race* while *deflationary on the spend level*—but this symmetry is not guaranteed. It may be that offensive cyber operations require expensive frontier-tier models while defense runs adequately on cheaper ones, or the reverse; tier asymmetry would break race-neutrality. We flag it as an explicit, unresolved caveat.

Update rule, pre-committed. At end-2027 and end-2028, the regime weights are re-scored against the realized signature statistics (via likelihoods specified in the parameter sheet), and the posterior weights are published next to the sealed prior. Forecast accuracy is evaluated by a proper scoring rule (CRPS or log score) against realized 2027–28 channel measurements, which scores the whole distribution rather than whether a point landed. Today's subjective weights become tomorrow's data-driven ones on a schedule—the registered-report spirit without the journal; formal registration, if pursued, is v2. For readers who want to fight the prior immediately, the published configuration ships with presets: Sealed (30/45/25), Flat (33/33/33), Offense-pessimist (50/35/15), and Defense-optimist (15/35/50), plus free sliders.

4.6 Endogenous safeguard response

The obvious objection—"new safeguards will be put in place; your model ignores the response"—was raised during sealing and is logged as registry D-15. The answer is that safeguards already enter the design in three places before any addition. (a) The excess-cost estimand is an *equilibrium* outcome: all deployed defense is netted out by construction, because the observables are what happened after defenders responded (D-02). (b) The defense side has explicit parameters—severity-truncation haircuts, containment probability, the sealed 20% patch-race odds, containment cycle time—and the regime mixture *is* the uncertainty over safeguard effectiveness. (c) Safeguards that succeed do not zero Δ; they shift cost from Channel 2 (losses) to Channel 1 (spend) and Channel 4 (friction), which is why all the channels are measured.

D-15 adds the fourth mechanism, the one dynamic the static design missed: safeguard *intensity responds to realized events*. If a simulated path's 2027 aggregate systemic loss exceeds a $50 billion trigger—roughly a Bashe-class realized year—the 2028 deviation-from-baseline of every AI multiplier and S3 probability on that path is damped by a factor with sealed prior 0.5/0.7/0.9. The anchor is the observed Hugging Face response: a near-miss with no recoverable dollar loss still produced, within weeks, an OpenAI research slowdown, agent-authored security fixes in deployment, and a UK AISI incident report. In the full run the trigger fires on 14.4% of 2027 paths. A matched-seed reduced-draw comparison (2,000×2,000, anchored variant, common random numbers; the trigger fires on 6.0% of its paths) isolates the effect: forcing damping off raises the 2028 mixture mean by $8 billion and the P99 by $56 billion while leaving the median essentially unchanged. That is the mechanism doing exactly what it should—realized 2027 systemic years pull 2028 back toward baseline, a negative autocorrelation that thins the two-year cumulative tail and touches only the tail. "Muddle through" is thereby partly *endogenous* in this model, not assumed, and now it has numbers.

4.7 Honesty constraints

Two constraints are stated up front because the P99 will be assumption-driven and a referee—or Tyler—should see that immediately. Twenty years of loss data containing a handful of systemic events means the pre-AI tail index already carries wide standard errors; and the AI shift is out-of-sample by construction—no historical data can validate the dependence-shift prior. The simulation is therefore a coherence engine, not an oracle: its value is forcing every assumption into a named parameter with a stated prior, propagating uncertainty honestly, and revealing—via global sensitivity analysis—which assumptions actually drive the answer. Our strong expectation, recorded before running: the tornado diagram will show the dependence parameter and the tail index dominating everything, with the frequency multipliers people argue loudest about barely moving the expected value. Publishing that, with code and priors, tells the discourse where the real uncertainty lives and what to go measure—and it preempts the false-precision critique, because we are not claiming to know the P99; we are showing exactly which beliefs produce which P99s.

5 Results

All numbers in this section are produced by the published simulation code from the sealed parameter sheet (Appendix A); placeholders below are typed to the output that fills them.

5.1 The predictive distribution

Table 1 reports the reference-specification predictive distribution of the AI-attributable delta Δ (equation 1): log-linear counterfactual, full 2015–2024 window, sealed regime weights 30/45/25. The headline rows pool both S3 variants with equal model weight (registry D-04), so the headline number integrates the Variant-A/Variant-B axis of model uncertainty; the per-variant rows are reported beneath it. Pooled means—and the pooled exceedance probabilities in Table 2—are exact equal-weight averages of the two full-run variants; pooled quantiles are computed from the published equal-weight pooled sample export, the same object from which the released site recomputes. Headline figures are transfer-inclusive; on the reference cell the resource/transfer split (D-09) is 60/40 in 2027 and 61/39 in 2028.

Table 1: Predictive distribution of the AI-attributable cost delta Δ, reference specification ($B/yr, global, transfer-inclusive). Pooled rows are the headline basis: both S3 variants with equal model weight (D-04).

Year	S3 basis	Median	Mean	P90	P99
2027	**pooled (headline)**	68.3	163.0	502.7	1,701
	anchored (Variant A)	61.9	133.7	465.9	1,174
	fresh (Variant B)	74.9	192.3	546.2	2,708
2028	**pooled (headline)**	115.5	261.6	879.2	2,407
	anchored (Variant A)	107.2	240.6	845.6	2,114
	fresh (Variant B)	122.9	282.6	914.0	2,778

Reference-cell figures are always quoted with their robustness range: across the nine counterfactual-matrix cells (three functional forms × three fitting windows), the 2028 pooled median spans $77 billion (quadratic × full window) to $116 billion (the reference cell itself)—$68 to $107 billion within the anchored variant alone—with the conservatism ordering running as pre-registered—the quadratic forms, which extrapolate pre-2025 acceleration, measure the least excess. The divergence between S3 Variant A (cat-model-anchored) and Variant B (agentic event tree) at the 2028 P99 is a factor of 1.31 (anchored $2,114 billion versus fresh $2,778 billion; the median ratio is 1.15, fresh median $123 billion, mean $283 billion), reported as model uncertainty per registry D-04: A imports pre-agentic propagation, B invents its own tail, and the gap between them is information.

5.2 Exceedance probabilities at the pre-committed thresholds

The four thresholds were committed before any code ran (registry D-06); they are round fractions of the ∼$1 trillion baseline and map onto recognizable positions in the debate.

5.3 The median–mean wedge

The headline finding is the shape, not a point. In the reference specification (pooled S3 basis) the 2028 mean ($262 billion) exceeds the 2028 median ($116 billion) by a factor of 2.3—a wedge of $146 billion—and the regime decomposition shows where the mean lives. The offense-dominant regime (pooled median $625 billion, mean $781 billion) contributes $234 billion of the $262 billion mixture mean at its 30% weight—roughly 90% of the expected value—while the mixture

Table 2: Exceedance probabilities at pre-committed thresholds, reference specification. Headline column is the pooled S3 basis; anchored / fresh variants in parentheses.

Threshold	$P(\Delta_{2027} > X)$	$P(\Delta_{2028} > X)$	Interpretation of the threshold
\$25B	63.6% (62.1 / 65.0)	66.3% (65.3 / 67.3)	≈ noise; Cowen effectively vindicated
\$50B	55.4% (53.7 / 57.1)	61.9% (60.8 / 63.0)	muddle-through validated
\$100B	42.5% (40.6 / 44.3)	52.5% (51.3 / 53.8)	material but manageable
\$200B	28.9% (27.1 / 30.6)	39.2% (38.0 / 40.5)	the tail matters

median sits near the balanced regime's \$98 billion and the defense-dominant regime is outright negative (median −\$148 billion). Years in which nothing NotPetya-class happens look like the median; the mean is an average over worlds most of which do not occur. This is the precise sense in which we think Cowen is right and still incomplete: right about the mode, which is unremarkable muddle-through; incomplete because "expected costs" is a mean, and the mean of a distribution this skewed is not a statement about typical years. The wedge is also exactly what the regime-weight knob moves. The mixture mean is linear in the weights, so from the published per-regime distributions the Flat preset (33/33/33) implies a 2028 pooled mean of \$257 billion, and Defense-optimist (15/35/50) compresses it to \$89 billion—readers can locate their own priors on that dial, and the released configuration recomputes the full distribution, not just the mean, under any weighting.

5.4 Sensitivity: the tornado

Table 3 reports the top eight of the sixty sealed parameters, ranked by the swing of the 2028 mixture mean (reference cell, anchored variant) when each parameter is pinned at its prior's P10 and P90; the baseline mean in these reduced-draw reruns (800×800) is \$234.6 billion. The full sixty-parameter diagram is rendered on the released site, and the complete table ships with the outputs (`out/tornado.csv`).

Table 3: Tornado: sensitivity of $E[\Delta_{2028}]$ to the sealed parameters, top 8 of 60 (\$B; mean with parameter pinned at prior P10/P90; baseline \$234.6B).

Parameter	Sealed name	Mean@P10	Mean@P90	Swi
S1 frequency multiplier, offense regime	`ai_shift.offense.s1_freq_mult`	143.0	348.2	205
S1 2024 frequency level	`strata.s1.freq_mean_2024`	157.7	331.8	174
S1 frequency multiplier, balanced regime	`ai_shift.balanced.s1_freq_mult`	187.9	290.3	102
S1 severity shift, offense regime	`ai_shift.offense.severity_shift`	187.8	282.9	95.
Reporting-propensity share of S1 growth	`strata.s1.reporting_growth_share`	277.4	191.4	85.
Defense severity truncation, balanced	`defense.balanced.severity_truncation`	264.3	198.8	65.
Defense severity truncation, offense	`defense.offense.severity_truncation`	259.7	207.6	52.
S1 severity shift, balanced regime	`ai_shift.balanced.severity_shift`	212.6	260.4	47.

Our pre-recorded expectation (Section 4) was that the S3 dependence parameter and the severity tail index would dominate, with the S1/S2 frequency multipliers far behind. On the pre-registered target—the 2028 *mean*—that expectation was simply wrong, and we say so. Volume-crime frequency assumptions dominate: the three largest swings are all S1 frequency parameters (\$205 billion for the offense-regime multiplier, \$174 billion for the 2024 frequency level, \$102 billion for the balanced-regime multiplier), and the reporting-propensity correction ranks fifth

at $86 billion. The S2 tail shape ξ moves the mean by only $17 billion, because with the D-14 caps sealed it is the *cap*, not ξ, that governs how much tail mass enters the mean—visible in the truncation haircuts at ranks six and seven. The S3 dependence parameters mattered far less to the mean than we expected: the offense-regime S3 probability multiplier swings $15 billion, and the dependence-correlation shifts move it by under $1 billion. Two honest glosses. First, the tornado's target is the mean, and parameters that act mainly at the far tail are structurally muted on that target; the released outputs do not include a P99 tornado, so we do not claim the dependence priors are unimportant at the P99—only that the argument about the *mean* turns out to be an argument about S1 frequency and reporting artifacts, not about exotic tail physics. Second, the ranking is a finding with an action attached: the 2027 measurement effort should prioritize the volume-crime frequency series (IC3 and its kin) and the reporting-propensity correction from the insured subsample, because those observables most reduce forecast variance.

5.5 The ILS-consistency check

Channel 3's calibration job culminates in one confrontation. The Matterhorn Re industry-loss-trigger cyber catastrophe bond attaches at $9 billion of US cyber industry insured loss with a market-implied annual attachment probability of about 2.2% (expected loss 1.72%, spread 12% at issue) (Artemis, 2023b); the first 144A cyber cat bond, Long Walk Re, priced at a 9.75% spread on a 1.97% expected loss (Artemis, 2023a); and the December 2025 PoleStar Re tranches price expected losses of 0.82–2.05% at spreads of 7.0–10.5% (Artemis, 2025). Mapping our simulated 2027 loss distribution through the sealed conversion priors (modal economic-to-insured ratio 7.1:1, US-to-global multiplier 2.45—together a global economic threshold of $156 billion) gives a model-implied probability of breaching the $9 billion industry trigger of **9.1%** under the anchored S3 variant and **15.1%** under the fresh variant, against the market's priced **2.23%**: a wedge of 4.1× to 6.8×.

The market-side context for this check is a finding in itself, and we frame the check with it rather than against it. As of August 10, 2026, *no repricing of cyber catastrophe bonds attributable to the Hugging Face incident is observable*: no spread widening or secondary-market movement has been reported; recent cyber cat-bond pricing has *declined* as the early novelty premium erodes (AM Best (via Artemis), 2026); cyber insurance rates were still softening through Q1 2026 after a cumulative ∼22–27% decline from the mid-2022 peak (Howden, 2025; Marsh, 2026); and reinsurance pricing fell at the January 2026 renewal (Guy Carpenter, 2025). Structural arguments have been offered for why cyber ILS should be insulated from AI-model-driven turbulence—per-occurrence triggers and high attachment points (Man Group (via Artemis), 2026). In short: **the market currently prices AI cyber risk as contained.** The check therefore asks a sharp question, and the answer is now on the table. At 4.1–6.8× the market-implied attachment probability, one of two things is true: our offense-weighted tail priors are wrong, or the market's priced attachment probabilities are too low—and both branches are informative. Both are questions of fact, not of positioning, and both are scheduled for adjudication by the 2027–28 re-scoring, which will grade our distribution and the market's against the same realized data. There is also a wedge already on the record that does not depend on our simulation at all: technical threat indicators (time-to-exploit, AI-phishing efficacy, agentic incident counts) have moved sharply while market prices have moved in the opposite direction. Someone is wrong. This paper is our sealed answer as to who, and the re-scoring will grade it.

5.6 The placebo panel

We committed to publishing this panel unconditionally—including if it embarrasses us. It does, a little, and in an instructive direction.

Channel 1 passes cleanly: measured spend excess over the placebo window 2025Q1–2026Q1 (out-of-sample, pre-break) is small and mixed-sign across the nine cells, from −$8.6 billion to +$6.1 billion per year against baselines of $210–225 billion. Channel 2 does not pass cleanly, and Table 4 shows the failure in full: measured loss excess is *negative in all nine cells*—the reference cell reads −$37.2 billion per year on a $296 billion baseline (−12.6%), and the nine cells span −$19 to −$61 billion. Each cell's 90% band individually covers zero, so no single cell is statistically distinguishable from a pass; but nine negative point estimates out of nine is a pattern, and we state its cause plainly rather than hiding behind the bands. Realized IC3 growth decelerated over 2022–25 (to roughly 21–33% per year, against spurts of up to +64% earlier in the series), so *every* pre-2025 trend family fitted on the real series overpredicts the placebo period. The quadratic cells overpredict most, because they extrapolate the earlier acceleration—this is the pre-registered D-12 conservatism ordering, now visible in data—and the structural cells, which include the placebo window in their fit by construction, are smallest.

Table 4: Placebo panel, Channel 2, window 2025Q1–2026Q1 ($B/yr). All nine cells negative. Structural cells are in-sample through 2026Q1 by construction.

Form	Window	Excess	Baseline	90% band	In-sample
log-linear	full 2015–2024 *(ref.)*	−37.2	296	[−149.0, 74.6]	no
log-linear	ex-COVID	−40.8	299	[−157.7, 76.2]	no
log-linear	short 2018–2024	−38.4	297	[−114.0, 37.1]	no
quadratic	full 2015–2024	−61.0	319	[−181.1, 59.2]	no
quadratic	ex-COVID	−52.3	311	[−175.1, 70.6]	no
quadratic	short 2018–2024	−38.0	296	[−115.0, 38.9]	no
structural	full 2015–2024	−22.8	281	[−127.3, 81.7]	yes
structural	ex-COVID	−23.4	282	[−131.5, 84.6]	yes
structural	short 2018–2024	−19.3	278	[−92.1, 53.5]	yes

Three consequences, stated in the order that matters.

First, the direction of the bias. A counterfactual family that overpredicts the placebo period will also tend to overpredict the 2027–28 baseline, and excess is measured *against* that baseline—so, to the extent the placebo miss carries forward, our measured AI-attributable excess is likely *understated*. The bias is conservative: it works against the paper's own thesis, deflating the headline delta rather than manufacturing it. A placebo failure in the other direction—positive excess where there should be none—would have been the damning one, because it would mean the machine fabricates AI costs out of trend misfit. This machine, on the evidence of its own falsification check, leans the other way.

Second, what we do about it: nothing, and that is the pre-commitment doing its job. Per the D-12 policy sealed before any code ran, the placebo panel is published as the falsification diagnostic and the parameters stay untouched. Only the trend family is on trial in this table—the sealed priors were never conditioned on the placebo window, and re-fitting or re-centering after seeing the result would convert a falsification check into a tuning loop.

Third, what a reader may do about it. A reader who wants a placebo-corrected headline can treat the placebo offset as an additive correction—mentally adding the reference cell's $37

billion to our 2027 pooled median of $68 billion, or applying any of the nine cell-level offsets in Table 4. We decline to make that adjustment ourselves, post hoc: it is a knob-adjacent judgment (which offset, which persistence assumption, which cells), and every knob in this project was to be sealed before results, not reached for after them. The honest summary is: the counterfactual family overpredicts the recent past, the headline is correspondingly conservative, and the full offset menu is published so nobody has to take our word for either claim.

6 What Would Change Our Minds

A sealed prior is only worth sealing if the conditions for abandoning it are stated in advance. Ours are as follows.

- **The placebo fails.** If measured excess over 2025Q1–2026Q1 is materially nonzero (Section 5.6), our counterfactuals are misfitted and the headline delta is suspect in the amount of the placebo excess. This check is available to every reader on publication—and, as Section 5.6 reports, it came back nonzero in the *conservative* direction: negative in all nine cells, meaning the counterfactual family overpredicts and the headline is understated rather than manufactured. The check ran, the result is published, and the parameters stayed sealed.

- **The agentic share shrinks.** The offense tilt rests on elicitation answer 1. If, by end-2028, incidents bearing agentic architectural signatures are a *shrinking* share of major incidents, the offense-dominant weight was too high and the posterior will show it.

- **The patch race is won.** We sealed a 20% probability that median enterprise patch latency beats exploit-development latency by 2028. If defenders close the gap—if agentic auto-remediation actually fixes the institutional deployment bottleneck at the median enterprise, not just the frontier—the defense-dominant regime deserves the weight, and the tail deflates accordingly. We consider this the most decision-relevant single observable in the entire design.

- **Channel 1 comes in below trend.** We sealed "above trend." Below-trend spend in 2028, absent a collapse in threat activity, is evidence of AI-driven unit-cost deflation of defense—the defense-dominant signature—and would move both the weights and the Channel-1 prior, which already straddles zero.

- **The market stays right.** If through 2028 cyber ILS spreads stay soft, insurance rates keep declining, no post-capability-event repricing ever materializes, and realized Δ_2 stays inside trend, then the market's "contained" verdict was correct, our tail priors were too fat, and the re-scoring will document by how much. Conversely, a Matterhorn-trigger breach or a disorderly repricing episode would update the other way.

- **Cowen's own criterion.** Realized Δ below $25 billion per year through 2028 is, per the pre-committed threshold table, "approximately noise": Cowen effectively vindicated, and we will say exactly that.

The mechanism is scheduled, not aspirational: at end-2027 and end-2028 the regime weights are re-scored against the realized signature statistics via the pre-specified likelihoods, the full

distribution is graded by CRPS/log score against realized channel measurements, and posterior sits next to sealed prior in public. Interim readouts arrive earlier on the fast channels—the options surface scores continuously, spend actuals by mid-2027, insured-loss data fully by 2028–29.

We took Cowen's post as an assignment, and this paper is the submission to its first two asks: a method that a skeptic can attack at named joints, and numbers that a scorekeeper can grade on a calendar. The disagreement that remains is narrow, quantified, and scheduled for resolution. That seems to us like acting as if it is science.

A Sealed Parameter Sheet

This appendix reproduces the sealed parameter sheet (v1, sealed August 10, 2026, author sign-off; D-14 ruling accepted, D-15 added on acceptance). *Convention:* priors are given as (p10 / mode / p90) unless noted; the engine maps these to lognormal or triangular hyperpriors in the outer epistemic loop. Every row runs anchor → transformation → prior. **[V]** marks an anchor verified to a primary source during the project's verification pass; **[U]** marks an unverified working value, used and flagged. *Estimand:* Δ = global, all-sector, transfer-inclusive AI-attributable excess cost versus pre-2025 trend, calendar 2027 and 2028. Rows marked ★ are the sheet's "argue-hardest" items—the most consequential judgments in the model, on which review time was deliberately concentrated (D-14's caps and ξ prior; the S3 dependence shift and correlation; the S1/S2 frequency multipliers; the reporting-propensity share; the category deflator; the Channel-1 labor multiplier).

A.0 Tail heaviness and severity caps (D-14) ★

The research came back heavier-tailed than the handoff's working value ($\xi \approx 0.6$–0.9, "near variance-infinite"): published dollar-loss fits imply tail index $\alpha \approx 0.6$–0.7, i.e. $\xi = 1/\alpha \approx 1.4$–$1.7$—infinite-*mean* territory—and Advisen-based work finds cyber loss distributions "by and large . . . of the infinite mean type" [V]; a minority lighter fit exists ($\xi \approx 0.20$). With $\xi \geq 1$, $E[\Delta]$ does not exist unless severities are truncated. Sealed resolution:

- **S2 per-event cap: $250B** (above the Cloud Down extreme $121B [V]; no single firm/event has approached it).
- **S3 annual aggregate cap: $3.5T/yr ≈ 3% of gross world product** (the Lloyd's $16T extreme is a five-year figure [V]; 3%/yr of GWP is war/pandemic-scale disruption—a declared plausibility ceiling, not a fact). Presets: Strict $1T/yr, Central $3.5T/yr (chosen), Permissive $10T/yr.
- ξ **prior (S2): 0.65 / 0.95 / 1.45**—spans the finite-variance sensitivity leg through the literature's modal infinite-mean finding. With the cap, all moments exist; the cap choice then partially determines the mean. This is arguably the single most consequential judgment in the model, alongside the dependence shift.

A.1 Stratum S1 — routine volume crime (BEC, fraud, commodity ransomware)

Parameter	Prior (p10/mode/p90)	Anchor → transformation
US reported losses, 2025 actual	$20.9B (point)	IC3 2025 report [V]
Underreporting multiplier	3 / 5 / 8	declared judgment; IC3 captures a minority of losses (FBI's own caveats); knob
US → global multiplier	2.0 / 2.4 / 3.0	US ≈ 35–45% of global digitized exposure; declared [U]
⇒ S1 global baseline level 2026	∼$150 / 260 / 450B	product of above, grown one year
Pre-2025 *measured* growth	25–35%/yr	IC3 2020–2025 series [V]
★ Reporting-propensity share of measured growth	0.15 / 0.35 / 0.55	share of measured growth that is reporting artifact, not incidence; anchored to insured subsample where reporting is contractual (Coalition/NetDiligence stable severity trends vs. IC3 steepness [V]); knob
⇒ True incidence + severity growth (baseline trend)	∼14 / 20 / 27%/yr	measured growth × (1 − reporting share)
Frequency dispersion (NB)	$k = 8$ / 15 / 30	IC3 YoY residual burstiness incl. 2021 spike treated stochastically (D-12)
Severity body (lognormal σ)	1.5 / 1.8 / 2.2	NetDiligence SME mean $264K, large-firm $10.3M; means ≫ medians [V]; μ set to match mean
Transfer share of S1	0.45 / 0.55 / 0.70	ransoms + fraud proceeds vs. resource costs (BI, remediation); Coalition/DBIR composition [V]; feeds D-09 split

A.2 Stratum S2 — major single-firm events (> $500M economic)

Parameter	Prior	Anchor → transformation
Baseline frequency (Poisson λ/yr, global)	1.5 / 3 / 6	realized ≥$500M events 2017–2024: Equifax, NotPetya components (Merck ∼$1.4B, Maersk ∼$300M), Change Healthcare $3.09B [V], MOVEit central ∼$15.8B modeled [V], others; count ambiguity declared
Baseline frequency trend	+3 / +8 / +15%/yr	severity-inflation in claims data [V]; declared
GPD threshold	$500M (fixed)	definition of stratum
★ GPD shape ξ	0.65 / 0.95 / 1.45	see A.0 (D-14) [V, units-corrected]
GPD scale	set s.t. median exceedance $1.2B (0.8–2.0)	anchors: Change Healthcare $3.09B, Merck $1.4B, MOVEit dispersion [V]
Per-event cap	$250B (knob, D-14)	Cloud Down extreme $121B [V] + headroom

A.3 Stratum S3 — systemic / correlated events

S3a — anchored variant (published scenarios, AI shift-factors applied).

Scenario	Annual prob (pre-AI)	Severity (economic, event total)	Anchor
Correlated infra failure (CrowdStrike-class+)	2 / 4 / 8%	\$8 / 25 / 120B	CrowdStrike \$5.4B F500-only ⇒ global ∼\$15B [V]; Cloud Down \$5.3–19B US 2018, ∼2× cloud-growth scaling ⇒ \$10–40B central [V; the old \$120B working bound was UNVERIFIED and is retired—now the p90+, not the range top]
Systemic ransomware/worm (Bashe-class)	0.5 / 1 / 2%	\$80 / 150 / 300B	CyRiM/Lloyd's Bashe \$85–193B [V], ∼1.3× economy scaling
Payments-infrastructure attack (Lloyd's 2023)	major 0.67%/yr @ \$2.2T total; severe 0.10%/yr @ ∼\$6T; extreme 0.024%/yr @ \$16T	first-year share 0.30 / 0.40 / 0.55 of total (declared—Lloyd's publishes no per-year path [V])	Lloyd's/CCRS triple used directly, NOT the \$3.5T weighted average [V]
Common-shock correlation ρ (pre-AI)	0.05 / 0.10 / 0.20	—	JRT Table B.XIII spillovers on unaffected peers 2–4× own-firm effects [V]—market already treats cyber as correlated

S3b — fresh agentic event tree (8 parameters, anchored to HF/Anthropic post-mortems).

#	Parameter	Prior	Anchor
1	Systemic-capable autonomous campaigns/yr, 2027	2 / 6 / 20	2026 observed: 1 HF + 3 Anthropic instances while capability nascent [V]; GTG-1002 shows state adoption (AI ran 80–90% of tactical ops) [V]
2	P(systemic foothold | campaign)	0.03 / 0.10 / 0.30	HF: sandbox escape + zero-day chain succeeded once-for-once observed; base-rate humility
3	Propagation branching factor R per containment cycle	0.4 / 1.2 / 3.0	HF: lateral movement + re-establishment 2 days post-patch [V]
4	Containment cycle time (days, lognormal median)	2 / 4 / 14	HF weekend timeline [V]
5	P(defender wins patch race)	0.10 / **0.20** / 0.40	**the author's sealed D-07 Q2 answer**, Mandiant TTE −7 days 2025 [V]
6	Blast-radius share of economy given monoculture	0.02 / 0.08 / 0.25	top-3 cloud ≈ 65% infra share; agent-framework concentration; declared
7	Damage rate during systemic event (\$B/day, full-economy-scaled)	4 / 10 / 30	CrowdStrike single-day \$5.4B F500-only ⇒ ∼\$15B global/day [V]
8	P(damaging objective | capable campaign)	0.2 / 0.5 / 0.8	HF was eval-seeking (no damage objective); criminal/state adoption trend [V]

A.4 AI shift-factors × regime ★

2027 multipliers on baselines; 2028 = 1.5× the deviation from 1.0, compounding drift. Regime mixture: **0.30 / 0.45 / 0.25 — SEALED (D-07)**.

Parameter	Offense-dom.	Balanced	Defense-dom.	Anchors
S1 frequency mult	1.3 / 1.6 / 2.2	1.0 / 1.15 / 1.4	0.8 / 0.95 / 1.1	AI-phishing click 54% vs 12% (4.5×) [V]; Hoxhunt +24% vs human red teams [V]; 67.3% of Anthropic-banned accounts used AI for malware [V]; counterweight: Mandiant "not yet the AI-breach year," ransomware payments flat-down 2024–25, payment rate 28% [V]
S1 severity shift	+10/+20/+35%	0/+5/+15%	−10/0/+5%	personalization scales fraud size; weak anchor, wide priors
S2 frequency mult	1.3 / 1.8 / 2.8	0.9 / 1.2 / 1.6	0.7 / 0.9 / 1.1	TTE negative [V]; vuln exploitation now #1 vector [V]
★ S3 dependence: prob mult on S3a scenarios	1.5 / 2.5 / 5.0	1.0 / 1.4 / 2.2	0.6 / 0.9 / 1.2	CyberCube: AI = "force-multiplier… aggregation and correlation risks" [V]; monoculture argument; HF propagation signature; **weakest-anchored, most consequential prior in the model**
★ S3 correlation ρ shift	$\rho \rightarrow$ 0.15/0.35/0.60	$\rho \rightarrow$ 0.10/0.20/0.35	$\rho \rightarrow$ 0.05/0.10/0.20	same as above
Defense truncation (severity haircut via faster containment)	0 / 0 / 10%	5 / 10 / 20%	15 / 25 / 40%	OpenAI agent-written patches [V-adjacent]; AI-SOC triage; anchored weakly, wide

A.5 Channel 1 — defensive expenditure above trend

Parameter	Prior	Anchor
Vendor-spend baseline 2026	$240B (232–248)	Gartner 2026F [V]; series breaks 2018/2024 handled by fitting growth rates, not levels
Baseline growth (pre-2025 trend)	11 / 12.5 / 14%/yr	Gartner/IDC concordant [V]
Internal labor cost multiplier (× vendor spend)	0.8 / 1.2 / 1.6	∼5.5M global workforce [V] × blended fully-loaded cost (declared [U]); WIDE—flagged
★ Category-inflation deflator	0.25 / 0.45 / 0.70	share of measured "AI-security" category spend that is genuinely incremental vs. relabeled; declared judgment; knob
Ch1 AI growth-rate delta (pp/yr, by regime)	off: +1/+3.5/+7 · bal: −1/+1/+3 · def: −5/−2/+1	symmetric per D-10; Black Hat spend signal [V-adjacent] vs. token-price deflation; tier-asymmetry caveat registered (D-07 Q3)

A.6 Channel 2 corrections (applied to S1+S2+S3 measured outcomes)

Parameter	Prior	Anchor
★ Post-break reporting-propensity uplift	5 / 12 / 25% apparent-loss inflation 2026–28	SEC 8-K migration to voluntary 8.01 (50 vs 29 filings [V]) + post-HF salience; corrected via insured subsample; knob
Transfer share (S1/S2/S3 blended)	0.35 / 0.45 / 0.60	component composition [V]; feeds D-09 headline split
Economic → insured ratio (for ILS check only)	5 / 7 / 10 : 1	CrowdStrike insured \$0.54–1.8B vs \$5.4B+ economic [V]

A.7 Channel 4 — trust tax, and the endogenous safeguard response (D-15)

Channel 4: triangular \$2B / \$12B / \$60B — **SEALED (D-13, Moderate)**; phase-in 50% weight 2027, 100% 2028. Sensitivity-only.

Parameter (D-15)	Prior	Anchor
Post-event damping factor (applied to 2028 deviation-from-baseline of all AI multipliers and S3 probabilities, if trigger hit in 2027)	0.5 / 0.7 / 0.9	observed Hugging Face response: near-miss with no recoverable loss still produced an OpenAI research slowdown, deployed agent-authored fixes, and a UK AISI incident report within weeks [V]
Damping trigger (2027 simulated aggregate systemic loss)	\$50B (knob)	roughly a Bashe-class realized year; declared

Interpretation (as sealed): safeguards already enter three ways—(a) the estimand is an equilibrium outcome, net of all deployed defense by construction; (b) explicit defense parameters (truncation haircuts, containment, patch-race odds); (c) successful safeguards shift cost from Ch2 to Ch1/Ch4 rather than zeroing Δ. D-15 adds the fourth: safeguard intensity responds to realized events, creating negative 2027→2028 autocorrelation and thinning the two-year cumulative tail. "Muddle through" is thereby partly endogenous, not assumed.

A.8 Channel 3 — validation targets (contribute \$0; pre-registered checks)

- Matterhorn Re attaches at **\$9B US industry insured loss**, attach prob 2.23% [V]: our simulated P(US insured loss $\geq$ \$9B) must be confrontable with it.
- PoleStar 2026-1 ELs 0.82–2.05%, spreads 7–10.5%, priced Dec 2025 *at/below guidance*; **no post-HF repricing observed as of Aug 10, 2026** [V]. The market currently prices AI-cyber as contained. If our offense-regime P(attach) $\gg$ market, we say so, report the wedge as a finding, and let the 2027–28 re-scoring adjudicate.
- Rate-on-line still softening through Q1 2026 (Marsh −5%) [V] while technical indicators (TTE, phishing efficacy) worsen—the wedge is itself a finding.

A.9 Placebo window (D-12)

Counterfactual applied 2025Q1–2026Q1 must show ≈0 excess. IC3 2025 grew +26%, inside the pre-break 25–35% trend [V]—passes on first inspection; formal check in the run. *(The formal check's result is reported in Section 5.6.)*

A.10 Honest-tension register (sealed verbatim)

Three cooling signals sit against the offense anchors: ransomware payments fell 35% in 2024 and stayed ∼flat in 2025 (payment-rate collapse to 28%); insurance rates softened through Q1 2026; Mandiant explicitly declines to call 2025 an AI-breach year. These discipline the balanced/defense priors and are reported, not buried.

B Decision Registry

Every methodological decision in this project was logged as it was made, with the alternatives considered and the rationale. Decisions marked **[KNOB]** are parameters a reader can change in the published configuration and re-run; decisions marked **[STRUCTURAL]** are design choices baked into the code architecture (still forkable, but not a slider). All decisions below are sealed as of August 10, 2026.

D-01. Core methodology — excess-cost design [STRUCTURAL]

Chosen: Excess-mortality template from epidemiology: fit pre-2025 baselines per channel, measure post-period deviations, attribute the AI share via signature statistics (time-to-exploit compression, social-engineering efficacy, agentic-architecture incident counts).
Alternatives: Adopt JRT's cross-sectional asset-pricing machinery wholesale; a pure structural offense/defense model.
Why: JRT's fixed effects discard the aggregate time-series variation the question is about; their $1.14T comes from a non-causal coefficient with an arbitrary counterfactual. Text-based indices are salience-contaminated post–Hugging Face.

D-02. Four channels, strict accounting [STRUCTURAL]

Chosen: $\Delta = \Delta_1$ (defensive spend) $+\Delta_2$ (realized losses) + bounded Δ_4 (friction/trust tax), each versus pre-2025 trend. Δ_3 (risk premia) contributes $0 to the headline—used for timing, validation, and tail calibration only.
Why: Valuation effects capitalize expected Channel 1+2 flows; adding them double counts.

D-03. Monte Carlo, two uncertainty layers [STRUCTURAL]

Chosen: Compound frequency–severity simulation, three event strata; outer loop draws parameters from priors (epistemic), inner loop simulates outcomes (aleatory); $\sim 10^4 \times 10^4$ draws. Full predictive distribution reported; the median–mean wedge is the headline finding.
Why: Heavy-tailed loss process (lognormal body + GPD tail, shape near the variance-infinite region); the annual aggregate is dominated by whether a NotPetya-class+ event occurs. Point estimates would be dishonest. *[Note added at severity verification: the published dollar-loss tail estimates are heavier than this log entry's phrasing—infinite-mean, not merely variance-infinite; see Section 4 for the two-leg parameterization adopted.]*

D-04. Stratum 3: both variants [STRUCTURAL]

Chosen: Run Variant A (anchored to published cat-model scenarios with AI shift-factors) *and* Variant B (fresh ∼8-parameter agentic event tree anchored to the Hugging Face/Anthropic postmortems). Report the divergence as model uncertainty.

Why: A borrows credibility but imports pre-agentic propagation assumptions; B is honest about the new threat model but invents its own tail. The divergence between them is itself information.

D-05. Counterfactual functional form [KNOB]

Chosen: Run all three forms per channel—log-linear, quadratic, structural (pre-committed break)—and report the full matrix. No single "primary" form hides the others.

Why: The functional form can manufacture or erase the entire AI delta; showing all three converts the biggest vulnerability into a transparency feature.

Knob: `counterfactual_form` ∈ {log_linear, quadratic, structural}.

D-06. Exceedance thresholds [KNOB]

Chosen: $P(\Delta > \$X)$ for $X \in \{\$25\text{B}, \$50\text{B}, \$100\text{B}, \$200\text{B}\}$, committed before any code ran.

Why: 2.5%/5%/10%/20% of the ∼$1T baseline; the thresholds map to debate positions: $25B ≈ noise, $50B ≈ muddle-through validated, $100B ≈ material but manageable, $200B ≈ the tail matters.

Knob: `thresholds`—array, user-editable.

D-07. Regime mixture weights [KNOB] — SEALED Aug 10, 2026

Chosen: 30% offense-dominant / 45% balanced / 25% defense-dominant—an asymmetric committed prior. Equal thirds rejected as partition-dependent pseudo-neutrality that dodges Cowen's ask for personal sealed beliefs. The tilt encodes one confident, falsifiable claim: diffusion speed favors offense on short horizons even if the technology is symmetric. The author's three sealed elicitation answers—(1) agentic-incident share by end-2028: *growing*; (2) probability the patch race is won by 2028: initial weakly held "yes," revised on the Mandiant time-to-exploit rebuttal to a sealed *20%*; (3) sign of Channel-1 deviation in 2028: *above trend*, with the token-price/tier-asymmetry caveat—are reproduced in full in Section 4.5.

Update rule (pre-committed): re-score the weights against realized 2027–28 signature statistics at end-2027 and end-2028; publish posterior next to sealed prior.

Knob: `regime_weights`—three sliders summing to 1. Presets: Sealed (30/45/25), Flat (33/33/33), Offense-pessimist (50/35/15), Defense-optimist (15/35/50).

D-08. Scope of Δ [STRUCTURAL]

Chosen: Global, all sectors (listed + private + government + household), with the listed-firm subset flagged as the cleanly measured core. Data sources with narrower scope (IC3 = US) are scaled with declared multipliers.

Alternatives: US-only; listed-firms-only (JRT's population).

Why: Cowen's question is about aggregate costs, not a sample. Scaling multipliers become declared priors rather than silent omissions.

D-09. Headline = transfer-inclusive [STRUCTURAL]

Chosen: The headline Δ includes transfers (ransoms, fraud proceeds); a split of resource

costs from transfers accompanies the headline everywhere it appears.
Why: Transfer-inclusive is what "costs to firms" means in Cowen's framing; the split prevents the conflation that produces $10T-style numbers.

D-10. Symmetric priors on Channel 1 [KNOB]
Chosen: The Channel-1 AI delta may be negative—AI may deflate the unit cost of defense (automated triage, agent-written patches), so spend can fall below trend. Priors straddle zero.
Why: The honest version; it makes the design a fair test rather than anxiety management. The published distribution may straddle zero, and that is acceptable.

D-11. Deliverable form [STRUCTURAL]
Chosen: A public online response to Cowen, not a peer-reviewed registered report. v1 = sealed dated priors; formal registration deferred to v2 if ever. Concrete deliverables: (a) this paper, as an Overleaf-ready LaTeX project; (b) the sealed parameter appendix and this decision registry, published; (c) code with parameters in a configuration JSON, all [KNOB] items exposed, outputs as JSON/CSV; (d) an interactive site with all knobs live (regime sliders and presets, Channel-4 posture presets, counterfactual-matrix selector, thresholds), the elicitation questions with a derived-weights flow, and this registry rendered with expandable rationale, distributions recomputed client-side; (e) the full conversation transcripts.
Why: Speed, legibility, forkability.

D-12. Baseline fitting window and reference specification [KNOB]
(a) Window menu—settled by construction: all three windows run and reported: full 2015–2024 (the 2021 ransomware spike treated stochastically—a tail draw informing the severity distribution, not excluded, not trend), ex-COVID (drop 2020–21 from trend fits), and short 2018–2024. Crossed with the three functional forms (D-05) = the 3×3 counterfactual matrix per channel. Key insight: for Channel 2 the Monte Carlo partly dissolves the window problem—the counterfactual is a fitted stochastic process, not a line, so spikes are observations of the tail, not anomalies. The window choice bites hardest on Channel 1 (spend trend), where COVID genuinely distorted 2020–21.
(b) Reference cell: log-linear $\times$ full window is the headline/abstract specification, always quoted with the min–max range across all nine cells in the same sentence. Conservatism ordering noted: forms that extrapolate a higher baseline measure less excess—roughly linear-in-levels (highest measured delta) > log-linear > log-quadratic (lowest, extrapolating pre-2025 acceleration continuing). Choosing the reference cell *is* choosing a conservatism level; log-linear is the middle of the dial.
(c) Structural break date: 2026Q2. Judgment: real cyber capability rose in 2026Q2 (the July 2026 Hugging Face incident disclosed capability that arrived shortly before). Alternatives considered: 2025Q1 (frontier-agentic-arrival framing)—rejected as too early relative to demonstrated capability; 2026Q3—rejected as a salience date, not a capability date. The structural form fits the baseline through 2026Q1 and breaks at 2026Q2.
Bonus falsification check (built into outputs): with baselines fitted on 2015–2024 and the break at 2026Q2, the quarters 2025Q1–2026Q1 form a placebo window—out-of-sample but pre-break, so measured excess there should be ≈ 0. Nonzero placebo excess = misfitted counterfactual, visible to every reader.

Knobs: `fit_window` ∈ {full_2015_2024, ex_covid, short_2018_2024}; `reference_cell` (form × window); `break_date`.

D-13. Channel 4 bounds [KNOB] — Moderate posture

Chosen: Triangular prior, min $2B / mode $12B / max $60B ("Moderate"). Sensitivity-only; no point estimate. Explicitly excluded even from the max: deterred-automation GDP effects (book territory, not a 2–3-year prediction).

Alternatives (presented in Section 3.2): Strict $1B/$5B/$25B—only what vendor filings can show (anchors: identity-verification vendor revenues, fraud-prevention headcount); Moderate $2B/$12B/$60B—proxy slice plus an early KYC/AML-analogy ramp (anchors: ∼$50B+/yr global KYC compliance as the mature-trust-tax benchmark; post-9/11 aviation security as the ramp template)—*chosen*; Expansive $5B/$25B/$100B—admits time-cost arithmetic (re-verification friction × wage rates), which gets large fast but is unfalsifiable.

Why Moderate: the min and mode stay tethered to measurable anchors; the max acknowledges the time-cost argument without endorsing its inputs and caps how much story can hide in the fuzzy channel.

Knob: `ch4_posture` ∈ {strict, moderate, expansive} as a preset, plus free sliders `ch4_min`/`ch4_mode`/`ch4_max`

D-14. Tail heaviness and severity caps [KNOB] — sealed Aug 10, 2026

Chosen: GPD shape prior $\xi = 0.65/0.95/1.45$; S2 per-event cap $250B; S3 annual aggregate cap $3.5T (≈3% of gross world product). The cap is presented as a first-class knob with guidance on choosing it—presets: Strict $1T/yr ("no annual cyber loss beyond a great-recession quarter"), Central $3.5T/yr (chosen; war/pandemic-scale ceiling), Permissive $10T/yr ("let the tail run"; near-uncapped, the mean becomes prior-dominated).

Trigger: anchor verification found the published literature heavier-tailed than the handoff's working value ($\xi \approx 0.6$–0.9, "near variance-infinite"): dollar-loss tail indices cluster at $\alpha \approx 0.6$–0.7, i.e., GPD shape $\xi \approx 1.4$–1.7—infinite-*mean* territory (Malavasi et al., 2022); the handoff's value was the conservative edge, likely a tail-index/GPD-shape units conflation. With $\xi \geq 1$ the simulated mean does not exist without truncation.

Why: with the caps, all moments exist—and the cap then partially determines $E[\Delta]$, which is why it must be a loud, named knob rather than an implementation detail. Readers who think our mean is too small or too large should turn this knob first (Section 4).

Knobs: `s2_cap_busd`, `s3_annual_cap_busd`, `xi_prior`.

D-15. Endogenous safeguard response [KNOB] — sealed Aug 10, 2026

Question raised at sealing: new safeguards will be put in place—how are they modeled?

Three mechanisms already in the design: (a) the excess-cost estimand is an equilibrium outcome—all deployed defense is netted out by construction (D-02); (b) explicit defense parameters: truncation haircuts, containment probability, patch-race odds (sealed 20%), containment cycle time—the regime mixture *is* the safeguard-effectiveness uncertainty; (c) safeguards that succeed do not zero Δ—they shift cost from Channel 2 (losses) to Channel 1 (spend) and Channel 4 (compliance/friction), which is why all channels are measured.

Chosen addition—post-event damping, the one dynamic not yet captured: safeguard intensity responds to realized events. If 2027's simulated aggregate systemic loss exceeds a trigger (default $50B), the 2028 deviation-from-baseline of all AI multipliers and S3 probabilities is damped by a factor with prior 0.5/0.7/0.9. Anchor: the observed Hugging Face response—a near-miss with no recoverable loss still produced, within weeks, an OpenAI

research slowdown, agent-authored security fixes in deployment, and a UK AISI incident report. Effect: negative autocorrelation between 2027 and 2028; thins the two-year cumulative tail; makes "muddle through" partly endogenous rather than assumed (Section 4.6 reports the realized effect).
Site requirement: a dedicated "Where is defense in this model?" explainer panel walking through mechanisms (a)–(c) and the D-15 damping loop, adjacent to the damping slider and trigger input—so visitors see that safeguards are priced in three places before they argue the model ignores them.
Knobs: `safeguard_damping` (prior), `damping_trigger_busd`.

References

AM Best (via Artemis). Pricing on recent cyber cat bonds has declined as the novelty premium diminishes. https://www.artemis.bm/news/topic/cyber-cat-bond/, April 2026.

Anthropic. Disrupting AI espionage. https://www.anthropic.com/news/disrupting-AI-espionage, 2026a.

Anthropic. AI-enabled cyber threats mapped to MITRE ATT&CK. https://www.anthropic.com/news/AI-enabled-cyber-threats-mitre-attack, 2026b.

Artemis. AXIS secures first Long Walk Re cyber cat bond at below-guidance pricing. https://www.artemis.bm/news/axis-secures-first-long-walk-re-cyber-cat-bond-at-below-guidance-pricing/, November 2023a.

Artemis. Swiss Re prices first industry-loss cyber cat bond (Matterhorn Re 2023-1). https://www.artemis.bm/news/swiss-re-prices-first-industry-loss-cyber-cat-bond/; deal terms: https://www.artemis.bm/deal-directory/matterhorn-re-ltd-series-2023-1/, December 2023b.

Artemis. Beazley secures the biggest cyber cat bond so far as PoleStar Re 2026-1 priced at $300m. https://www.artemis.bm/news/beazley-secures-the-biggest-cyber-cat-bond-so-far-as-polestar-re-2026-1-priced-at-300m/, December 2025.

Artemis. Catastrophe bond and ILS market report, Q4 2025. Technical report, Artemis.bm, 2026. https://www.artemis.bm/wp-content/uploads/2026/01/catastrophe-bond-ils-market-report-q4-2025.pdf.

Axios. White house confirms NotPetya caused $10 billion in damages. https://www.axios.com/2018/02/15/white-house-confirms-notpetya-1518728781, 2018.

Axios. OpenAI details Hugging Face breach at Black Hat. https://www.axios.com/2026/08/06/openai-hugging-face-black-hat, August 2026.

Antoine Bouveret. Cyber risk for the financial sector: A framework for quantitative assessment. Technical Report WP/18/143, International Monetary Fund, 2018.

Chainalysis. Ransomware payments in 2023 exceeded $1 billion. https://www.chainalysis.com/blog/ransomware-2024/, 2024.

Chainalysis. Crypto crime: Ransomware and victim extortion, 2025 update. https://www.chainalysis.com/blog/crypto-crime-ransomware-victim-extortion-2025/; see also https://cyberscoop.com/ransomware-payments-drop-35-percent-2024-chainalysis/, 2025.

Yuyu Chen, Paul Embrechts, and Ruodu Wang. Infinite-mean models in risk management. arXiv:2408.08678, https://arxiv.org/pdf/2408.08678, 2024.

Todd E. Clark and Kenneth D. West. Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1):291–311, 2007.

Cloud Security Alliance. Hugging Face CISO post-mortem. https://cloudsecurityalliance.org/artifacts/hugging-face-ciso-post-mortem, 2026.

CNBC. OpenAI models hack Hugging Face in sandbox-escape incident. https://www.cnbc.com/2026/07/22/open-ai-cyber-models-hack-hugging-face.html, July 2026.

Coalition. 2025 cyber claims report. https://www.coalitioninc.com/blog/cyber-insurance/2025-cyber-claims-report, 2025.

Roger M. Cooke. *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford University Press, 1991.

Tyler Cowen. Act like it is science. Marginal Revolution (blog), https://marginalrevolution.com, August 2026.

CPO Magazine. Mandiant: 70% of exploited vulnerabilities in 2023 were zero-days. https://www.cpomagazine.com/cyber-security/mandiant-70-of-exploited-vulnerabilities-in-2023-were-zero-days/, 2024.

CyberCube. Cybercube estimates global insured losses from the CrowdStrike event. https://www.cybcube.com/news/cybercube-estimates-global-insured-losses-from-crowdstrike-event, 2024a.

CyberCube. Projecting cyber insurance growth: Capital requirements at a 1-in-250 aggregate loss. https://www.businesswire.com/news/home/20240904059795/en/, September 2024b.

CyberCube. AI treated as force-multiplier for cyber losses; introduces aggregation and correlation risks. https://www.artemis.bm/news/ai-treated-as-force-multiplier-for-cyber-losses-introduces-aggregation-correlation-risks-cybercube/, April 2026.

CyberScoop. FBI IC3 annual cybercrime report: $20.9 billion in reported losses for 2025. https://cyberscoop.com/fbi-internet-crime-complaint-center-annual-cybercrime-report/, April 2026.

Cybersecurity Dive. Zero-day exploits in Google's annual report: Enterprise technologies increasingly targeted. https://www.cybersecuritydive.com/news/zero-day-exploits-google-report-vulnerabilities-enterprise/746556/, 2026.

Debevoise & Plimpton. Cybersecurity incident disclosure form 8-k tracker: Two-year update. https://www.debevoisedatablog.com/2026/05/21/cybersecurity-incident-disclosure-form-8-k-tracker-two-year-update/, May 2026.

Paul Dreyer, Therese Jones, Kelly Klima, Jenny Oberholtzer, Aaron Strong, Jonathan W. Welburn, and Zev Winkelman. Estimating the global cost of cyber risk: Methodology and examples. Technical Report RR-2299, RAND Corporation, 2018. https://www.rand.org/pubs/research_reports/RR2299.html.

Benjamin Edwards, Steven Hofmeyr, and Stephanie Forrest. Hype and heavy tails: A closer look at data breaches. *Journal of Cybersecurity*, 2(1):3–14, 2016. https://academic.oup.com/cybersecurity/article/2/1/3/2736315.

Martin Eling and Jan Wirfs. What are the actual costs of cyber risk events? *European Journal of Operational Research*, 272(3):1109–1119, 2019.

Emsisoft. Unpacking the MOVEit breach: Statistics and analysis. https://www.emsisoft.com/en/blog/44123/unpacking-the-moveit-breach-statistics-and-analysis/; see also https://techcrunch.com/2023/08/25/moveit-mass-hack-by-the-numbers/, 2023.

FBI Internet Crime Complaint Center. 2024 internet crime report. Technical report, Federal Bureau of Investigation, 2025. https://www.ic3.gov/AnnualReport/Reports/2024_IC3Report.pdf.

FBI Internet Crime Complaint Center. 2025 internet crime report. Technical report, Federal Bureau of Investigation, 2026. https://www.ic3.gov/AnnualReport/Reports/2025_IC3Report.pdf; reports index: https://www.ic3.gov/annualreport/reports.

Fortinet. Fortinet reports fourth quarter and full year 2025 financial results. https://www.fortinet.com/corporate/about-us/newsroom/press-releases/2026/fortinet-reports-fourth-quarter-full-year-2025-financial-results, 2026.

Fortune. OpenAI–Hugging Face hack: New details — everything we know and don't know. https://fortune.com/2026/07/29/openai-hugging-face-new-details-hack-everything-we-know-dont-know/, July 2026.

Gartner. Gartner forecasts worldwide information security spending to exceed $124 billion in 2019. https://www.gartner.com/en/newsroom/press-releases/2018-08-15-gartner-forecasts-worldwide-information-security-spending-to-exceed-124-billion-in-2019, 2018.

Gartner. Gartner forecasts global security and risk management spending to grow 14 percent in 2024. https://www.gartner.com/en/newsroom/press-releases/2023-09-28-gartner-forecasts-global-security-and-risk-management-spending-to-grow-14-percent-in-2024, 2023.

Gartner. Gartner forecasts worldwide end-user spending on information security to total $213 billion in 2025. https://www.gartner.com/en/newsroom/press-releases/2025-07-29-gartner-forecasts-worldwide-end-user-spending-on-information-security-to-total-213-billion-us-dollars-in-2025, July 2025.

Google Threat Intelligence Group. 2024 zero-day exploitation trends. https://cloud.google.com/blog/topics/threat-intelligence/2024-zero-day-trends/, 2025.

Greenberg Traurig. SEC cybersecurity disclosure trends: 2025 update on corporate reporting practices. https://www.gtlaw.com/en/insights/2025/2/sec-cybersecurity-disclosure-trends-2025-update-on-corporate-reporting-practices, 2025.

Guy Carpenter. Cyber reinsurance pricing under pressure at 1/1/2026 renewal. https://www.theinsurer.com/cyber-risk/news/guy-carpenter-cyber-reinsurance-pricing-under-pressure-at-11-while-more-tailored-2025-12-29/; see also https://www.reinsurancene.ws/non-proportional-cyber-pricing-falls-up-to-25-as-xol-and-hard-retro-gain-ground-at-1-1-guy-carpenter/, December 2025.

Tarek A. Hassan, Stephan Hollander, Laurence van Lent, and Ahmed Tahoun. Firm-level political risk: Measurement and effects. *Quarterly Journal of Economics*, 134(4):2135–2202, 2019.

Help Net Security. IDC: Cybersecurity spending to grow 12.2% in 2025, $377 billion by 2028. https://www.helpnetsecurity.com/2025/03/28/idc-cybersecurity-spending-2025/, 2025.

Howden. Rebooting growth: Global cyber insurance market report 2025. Technical report, Howden Group, 2025.

Hoxhunt. AI-generated spear phishing versus elite human red teams: A 70,000-simulation study. Hoxhunt research publication, 2025.

Hugging Face. Security incident — July 2026. https://huggingface.co/blog/security-incident-july-2026, July 2026a.

Hugging Face. Agent intrusion: Technical timeline. https://huggingface.co/blog/agent-intrusion-technical-timeline, July 2026b.

IDC. Global security spend to exceed $300 billion in 2026 as adoption of AI-driven security platforms gains momentum. https://www.biztechreports.com/news-archive/2026/3/20/global-security-spend-to-exceed-300-billion-in-2026-as-the-adoption-of-ai-driven-security-platforms-gains-momentum-idc-march-23-2026, March 2026.

ISC2. 2024 cybersecurity workforce study. https://www.isc2.org/Insights/2024/10/ISC2-2024-Cybersecurity-Workforce-Study, 2024.

Rustam Jamilov, Hélène Rey, and Ahmed Tahoun. The anatomy of cyber risk. *Journal of Finance*, July 2026. Firm-level exposure data: https://doi.org/10.5281/zenodo.21383844.

Edwin T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

Lloyd's of London. Lloyd's systemic risk scenario reveals global economy exposed to $3.5trn from major cyber attack. https://www.lloyds.com/insights/media-centre/press-releases/lloyds-systemic-risk-scenario-reveals-global-economy-exposed-to-3.5trn-from-major-cyber-attack; coverage: https://www.insurancejournal.com/news/international/2023/10/18/744702.htm, October 2023.

Lloyd's of London and AIR Worldwide. Cloud down: Impacts on the US economy. Technical report, Lloyd's Emerging Risk Report, 2018. https://assets.lloyds.com/assets/pdf-air-cyber-lloyds-public-2018-final/1/pdf-air-cyber-lloyds-public-2018-final.pdf; coverage: https://www.insurancejournal.com/news/national/2018/01/23/478138.htm.

Lloyd's of London and Cyence. Cloud down: Counting the cost of cloud provider failure. https://www.lloyds.com/clouddown; coverage: https://www.cnbc.com/2017/07/17/global-cyberattack-could-spur-53-billion-in-losses-lloyds-of-london.html, 2017.

Matteo Malavasi, Gareth W. Peters, Pavel V. Shevchenko, Stefan Trück, Jiwook Jang, and Georgy Sofronov. Cyber risk frequency, severity and insurance viability. arXiv:2111.03366, https://arxiv.org/pdf/2111.03366, 2022.

Man Group (via Artemis). Per-occurrence structures and high attachment points insulate cyber cat bonds from AI-model-driven turbulence. https://www.artemis.bm/news/topic/cyber-cat-bond/, July 2026.

Mandiant / Google Cloud. Time-to-exploit trends 2023: How low can you go? https://cloud.google.com/blog/topics/threat-intelligence/time-to-exploit-trends-2023, 2024.

Mandiant / Google Cloud. M-Trends 2026. https://cloud.google.com/blog/topics/threat-intelligence/m-trends-2026/; see also https://www.helpnetsecurity.com/2026/03/24/mandiant-m-trends-2026-report/, 2026.

Marsh. Global insurance market index: Cyber pricing declines continue through Q1 2026. https://www.marsh.com/en/services/international-placement-services/insights/global-insurance-market-index.html; see also https://www.marsh.com/en/services/cyber-risk/insights/cyber-insurance-market-update.html, 2026.

Microsoft. Microsoft digital defense report 2025. Technical report, Microsoft Corporation, 2025.

Munich Re. Cyber insurance: Risks and trends 2025. https://www.munichre.com/en/insights/cyber/cyber-insurance-risks-and-trends-2025.html, 2025.

NetDiligence. Cyber claims study 2025. https://cyberinsurancenews.org/cyber-insurance-claims-2025-netdiligence/; 2024 edition: https://netdiligence.com/wp-content/uploads/2025/04/NetDiligence-Cyber-Claims-Study-2024-Report-V1.1.pdf, 2025.

Okta. Okta announces fourth quarter fiscal year 2026 financial results. https://www.okta.com/newsroom/press-releases/okta-announces-fourth-quarter-fiscal-year-2026-financial-results/, 2026.

OpenAI. Hugging Face model evaluation security incident. https://openai.com/index/hugging-face-model-evaluation-security-incident/, July 2026.

Parametrix. CrowdStrike to cost Fortune 500 $5.4 billion; insured loss range of $540 million to $1.08 billion. https://www.parametrixinsurance.com/in-the-news/crowdstrike-to-cost-fortune-500-5-4-billion-insured-loss-range-of-540-million-to-1-08-billion, 2024.

Zacharias Sautner, Laurence van Lent, Grigory Vilkov, and Ruishen Zhang. Firm-level climate change exposure. *Journal of Finance*, 78(3):1449–1498, 2023.

Philip E. Tetlock and Dan Gardner. *Superforecasting: The Art and Science of Prediction*. Crown, 2015.

The Record. Ransomware payments tracked by Chainalysis in 2025. https://therecord.media/ransomware-payments-chainalysis-cybercrime, February 2026.

TIME. Inside the OpenAI–Hugging Face attack. https://time.com/article/2026/07/24/openai-hugging-face-attack/, July 2026.

UK AI Security Institute. Incident report: Unsanctioned agent behaviour during cyber testing. https://www.aisi.gov.uk/blog/incident-report-unsanctioned-agent-behaviour-during-cyber-testing, 2026.

UnitedHealth Group. Form 10-k, fiscal year 2024. https://www.sec.gov/Archives/edgar/data/731766/000073176625000063/unh-20241231.htm, 2025.

U.S. Bureau of Labor Statistics. Occupational outlook handbook: Information security analysts. https://www.bls.gov/ooh/computer-and-information-technology/information-security-analysts.htm, 2026.

Verizon. 2026 data breach investigations report. Technical report, Verizon Business, 2026. https://www.verizon.com/business/resources/Td15/reports/2026-dbir-data-breach-investigations-report.pdf; summary: https://www.helpnetsecurity.com/2026/05/20/verizon-2026-dbir-findings/.

Bennet von Skarczinski, Mathias Raschke, and Frank Teuteberg. Modelling maximum cyber incident losses of German organisations: An empirical study and modified extreme value distribution approach. *Geneva Papers on Risk and Insurance — Issues and Practice*, 48(2): 463–501, 2023. https://pmc.ncbi.nlm.nih.gov/articles/PMC10100641/.

Spencer Wheatley, Annette Hofmann, and Didier Sornette. Data breaches in the catastrophe framework and beyond. arXiv:1901.00699, https://arxiv.org/pdf/1901.00699, 2019.

Zscaler. Zscaler reports fourth quarter and fiscal 2025 financial results. https://ir.zscaler.com/news-releases/news-release-details/zscaler-reports-fourth-quarter-and-fiscal-2025-financial-results, 2025.