

A DATA-DRIVEN CISO BLUEPRINT

QUANTIFYING THE **HUMAN** ATTACK SURFACE

The science of adversary-realistic phishing simulation — and how to turn a click rate into a dollar-denominated, board-ready measure of human risk.



Published by Cybears · Your Cyber Resiliency Partner

xphish.cybears.net

VERSION 1.0

ABSTRACT

A CLICK RATE IS NOT A RISK METRIC

Phishing remains the dominant initial-access vector into the enterprise, and the rise of generative AI has made lures cheaper, faster, and more convincing than at any point in history. Yet the way most organizations test their defenses against it has barely evolved: an annual or quarterly phishing test, a single click-rate number, and no statistically defensible way to say whether human risk is rising or falling — let alone what it is worth in dollars.

This whitepaper argues that **phishing simulation must be treated as a measurement instrument, not a compliance ritual**. Done properly — adversary-realistic, continuous, multi-channel, and statistically rigorous — it produces something no other control can: a quantified, trended estimate of the organization's human attack surface, expressed in the language the board already uses to make decisions.

Drawing on industry breach research, behavioral science, and measurement theory, we examine why single-shot testing is statistically meaningless, define the properties of a credible simulation program, and present a model for converting behavioral signals into **financial exposure**. We then detail how the **Cybeards X Phish** platform operationalizes that model — AI-generated omnichannel attacks, adaptive difficulty, real-time teachable moments, and a dollar-denominated Human Risk Score — and back it with longitudinal and cohort evidence, a deployment blueprint, and a compliance mapping.

Who should read this

CISOs and security leaders who need to quantify and reduce human risk; red-team and security-operations leads designing realistic test programs; and GRC teams who must evidence effectiveness, not just activity, to auditors and the board.

CONTENTS

WHAT'S INSIDE

00	Executive Summary	04
01	The Modern Phishing Threat Volume, economics, channel mix, and the generative-AI escalation — in data.	06
02	Why Single-Shot Testing Fails Statistical noise, click-rate decay, sample size, and negative training.	12
03	Adversary-Realistic Simulation Threat emulation, the phishing kill chain, and omnichannel testing.	17
04	The Cybears X Phish Engine Closed-loop architecture, AI attack generation, targeting, and automation.	23
05	Quantifying Risk in Dollars The Human Risk Score, the loss-exposure model, and benchmarking.	29
06	Evidence & Outcomes Longitudinal results, a cohort study, risk distribution, and a deployment.	35
07	Deployment, Sovereignty & Compliance Deployment models, MEA-native residency, frameworks, and MSSP scale.	41
08	Conclusion & References The path forward, methodology, and sources.	45

14

data visualizations

8

chapters

47

pages

EXECUTIVE SUMMARY

TURN THE CLICK RATE INTO A NUMBER THE BOARD CAN ACT ON

Phishing is not a solved problem; it is an accelerating one. The defenders' core difficulty is no longer awareness that the threat exists — it is the absence of a credible, continuous measurement of how exposed the organization actually is. A single annual test cannot supply it, and a bare click-rate percentage tells the board nothing about likelihood, impact, or money.

The numbers that frame the problem

3.4B

phishing emails sent every day worldwide.

\$2.9B

annual reported losses to business email compromise.

<60s

median time from delivery to first click on a lure.

4.7*

rise in AI-crafted phishing since generative tools went mainstream.

The thesis

A phishing simulation program should be engineered as a **measurement instrument**. When it is adversary-realistic, continuous, multi-channel, and statistically sound, it yields a trended, dollar-denominated estimate of human risk — the one figure that lets a CISO answer the board's real question: **how much exposure do our people represent, and is it falling?**

"You cannot manage what you measure badly. A click rate from one annual test is measured badly — it is a single noisy sample of a moving target."

SIX PROPERTIES OF A CREDIBLE PROGRAM

The programs that actually reduce — and quantify — human risk share six properties. Each marks a departure from the annual-test status quo, and each becomes a concrete capability in the chapters that follow.

01 Continuous

Many small tests over time beat one large test a year — statistically and pedagogically.

02 Adversary-realistic

Lures that mirror live attacker tradecraft, not the obvious template everyone already spots.

03 Omnichannel

Email, SMS, voice, and QR — because attackers no longer confine themselves to the inbox.

04 Statistically sound

Sample sizes and trends that separate real signal from random month-to-month noise.

05 Financially expressed

Behavior translated into expected loss, so risk competes for budget on equal terms.

06 Closed-loop

Every result drives the next test and an automatic teachable moment — test feeds training feeds test.



The pages ahead supply the evidence for each property and show how **Cybeards X Phish** turns them into one running instrument — producing a Human Risk Score the board reads as exposure, and a trend line the CISO reads as progress.

01

THE MODERN PHISHING THREAT

Four data points define the battlefield: how much phishing there is, what it costs, where it now lands, and how generative AI changed the economics overnight.

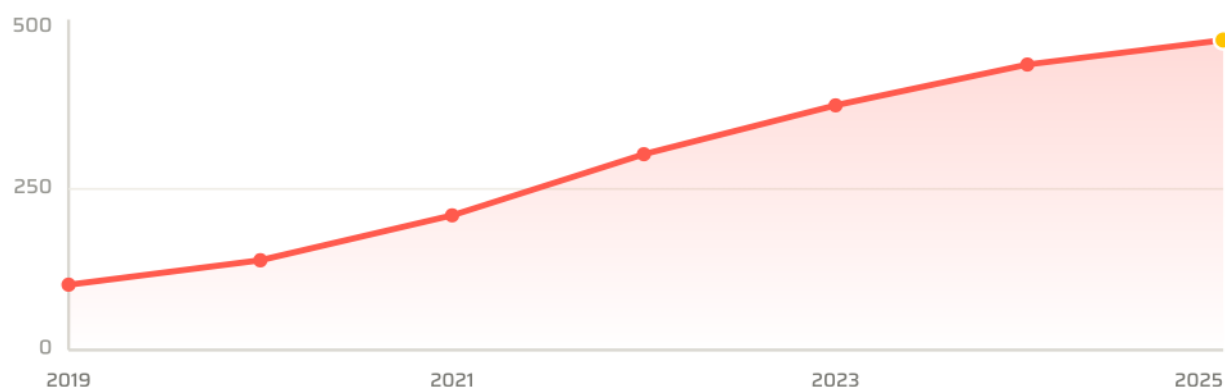
1.1 · VOLUME & GROWTH

THE MOST INDUSTRIALIZED ATTACK ON EARTH

Phishing is not a niche technique; it is the highest-volume attack category in existence. Billions of malicious messages are dispatched daily, and the trend line points in one direction. As detection improves, attackers simply send more — the economics reward volume, because a campaign needs only a single success to pay for itself thousands of times over.

Reported phishing & impersonation attacks, indexed

Relative volume, 2019 = 100 · indicative of published industry trend

**3.4B**

phishing emails dispatched every day across the internet.

~1 in 3

untrained users will click a well-crafted lure on first exposure.

5×

growth in indexed attack volume across the period shown.

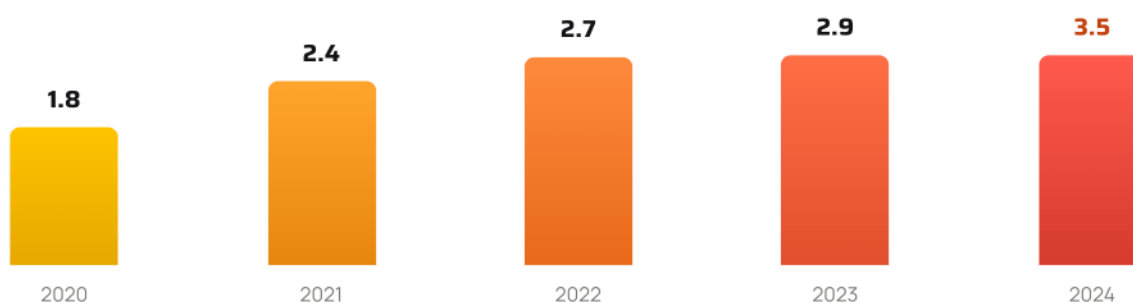
1.2 · THE ECONOMICS

WHERE A SINGLE CLICK ENDS UP COSTING MILLIONS

The damage from phishing concentrates in business email compromise (BEC) — the quiet, high-value fraud that follows a successful credential or trust compromise. Unlike ransomware, BEC rarely makes headlines, but its reported losses dwarf most other categories and have climbed every year that records have been kept.

Reported BEC losses by year

Annual reported losses, US\$ billions · indicative of published figures



Why the attacker math is so favorable

COST TO ATTACK

Pennies

Kits, domains, and AI-written copy are effectively free at scale.

PAYOUT PER HIT

\$120K+

Average wire fraud from a single successful BEC.

BREAK-EVEN

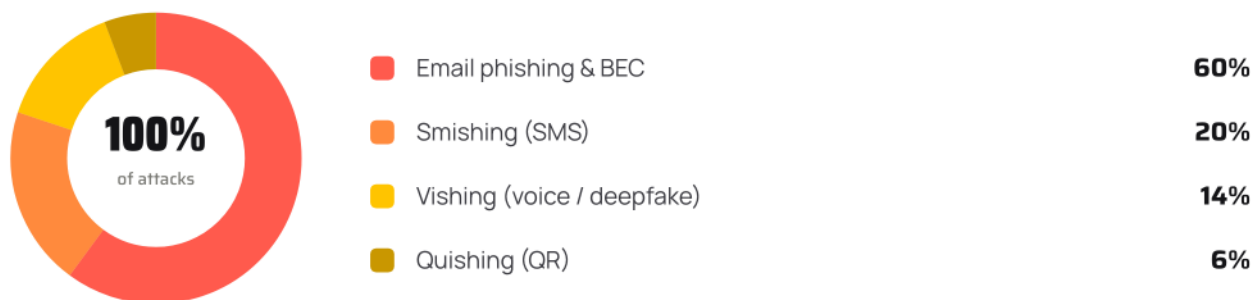
1 click

One success in a million messages funds the next campaign.

1.3 · CHANNEL MIX

THE ATTACK LEFT THE INBOX

For years "phishing" meant email, and defenses were built accordingly. Attackers responded by diversifying. Smishing (SMS), vishing (voice, increasingly with cloned audio), and quishing (malicious QR codes) now carry a fast-growing share of attacks precisely because they bypass email security gateways and catch users off guard on channels they never associated with phishing.



The strategic point is not the exact split — it shifts constantly — but the principle: **a test program that only probes email measures a shrinking fraction of the real attack surface.** The fastest-growing vectors are exactly the ones most organizations never simulate, which means their true exposure is systematically underestimated. Closing that gap requires testing every channel an attacker can reach, the subject of Chapter 3.

Quishing is the clearest example: a QR code in an email image sails past link scanners, then moves the victim onto a personal phone outside every corporate control.

1.4 · THE GENERATIVE-AI ESCALATION

THE TELLS ARE GONE, THE VOLUME IS UP

Generative AI changed phishing on two axes at once. It removed the quality ceiling — every lure is now fluent, contextual, and brand-perfect — and it removed the cost floor, so attackers can personalize at a scale that was previously impossible. The advice a generation of employees was given ("look for bad grammar") is now actively misleading.

Effect of AI-crafted lures on user click rate

Relative click rate vs. generic template · indicative of red-team findings

**4.7***

rise in AI-attributable phishing since mainstream LLMs.

***0**

grammar & spelling tells left to teach employees.

seconds

to generate a flawless, personalized, multi-language lure.

The implication for defenders is unforgiving: a test program must keep pace with attacker tradecraft, which now evolves at machine speed. Only a simulation engine that itself uses AI to generate fresh, realistic lures can credibly measure exposure to the attacks employees will actually face — the design principle behind Chapter 4.

1.5 · A STUDY WORTH NOTING

WHAT LARGE-SCALE TESTING REVEALS

Across published large-scale phishing studies — spanning hundreds of thousands of users — a consistent pattern emerges that should reshape how leaders think about human risk. Three findings recur often enough to be treated as planning assumptions.

A**Risk is concentrated, not uniform**

A small fraction of users — often under 10% — account for the majority of repeat clicks. Targeting them yields disproportionate risk reduction.

B**Reporting is the leading indicator**

A rising report rate predicts falling real-world compromise more reliably than a falling click rate alone.

C**Difficulty matters more than frequency**

Realistic, varied lures build durable resilience; repeated easy tests inflate scores without changing behavior under real pressure.

CHAPTER 1 · KEY TAKEAWAYS

- ◆ Phishing is the highest-volume attack category and still growing; defenders cannot out-filter it.
- ◆ BEC concentrates the financial damage, and the attacker economics overwhelmingly favor volume.
- ◆ Attacks have spread beyond email; email-only testing understates true exposure.
- ◆ Generative AI raised lure quality and volume together, so testing must evolve at machine speed.

02

WHY SINGLE-SHOT TESTING FAILS

An annual phishing test is not just weak training — it is a statistically invalid measurement. This chapter shows why, with the numbers.

2.1 · THE SINGLE-SAMPLE PROBLEM

ONE TEST IS ONE NOISY DATA POINT

A phishing click rate is a measurement, and every measurement carries uncertainty. A single annual test produces one observation, with a confidence interval wide enough to swallow most of the changes a CISO would care about. Two consecutive years can differ by several points for reasons that have nothing to do with employee behavior — the lure's theme, the send time, the season, who happened to be on holiday.

Measured click rate with 95% confidence interval

One annual test (wide band) vs. twelve monthly tests (tight band) · illustrative



The statistics are not subtle. Spreading the same population across many smaller, frequent tests shrinks the confidence interval dramatically — turning a number you cannot trust into a trend you can. **The annual test does not just teach badly; it measures badly**, reporting movements that are statistically indistinguishable from noise as if they were progress or regression.

A CISO who reports "click rate fell from 18% to 14%" off two annual tests may be reporting pure chance. The same four-point move, sustained across twelve monthly tests, is a real and defensible result.

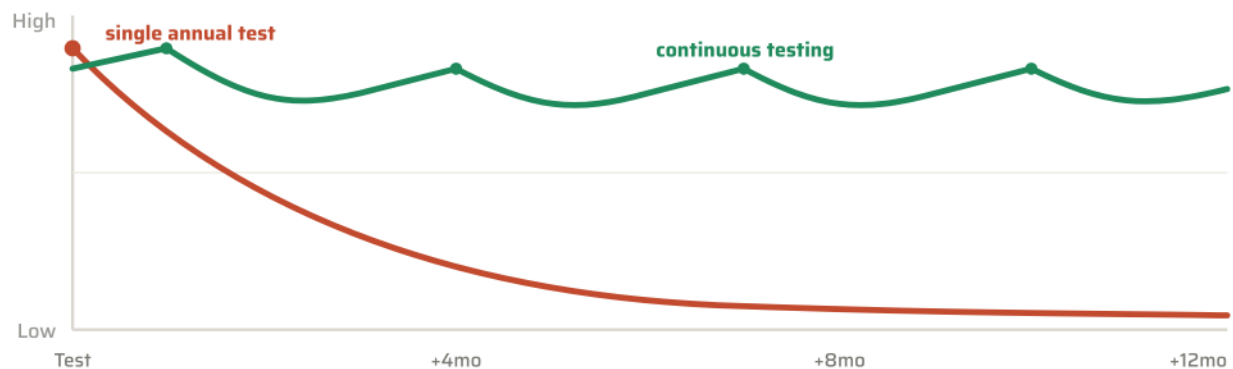
2.2 · RESILIENCE DECAY

WHATEVER A TEST TEACHES, IT FADES

Even when a single test does change behavior, the effect is temporary. Resilience built in one sitting decays much like any other learned skill: sharp immediately after, eroding week by week, and largely gone within a quarter. An annual cadence therefore leaves the workforce vulnerable for the overwhelming majority of the year — protected briefly after the test, exposed for the long months until the next one.

Resilience after a single test vs. continuous testing

Relative resilience over 12 months · illustrative



~90 days

before single-test resilience largely fades.

9 / 12

months a year the annual-only workforce sits unprotected.

Sustained

resilience only continuous testing maintains.

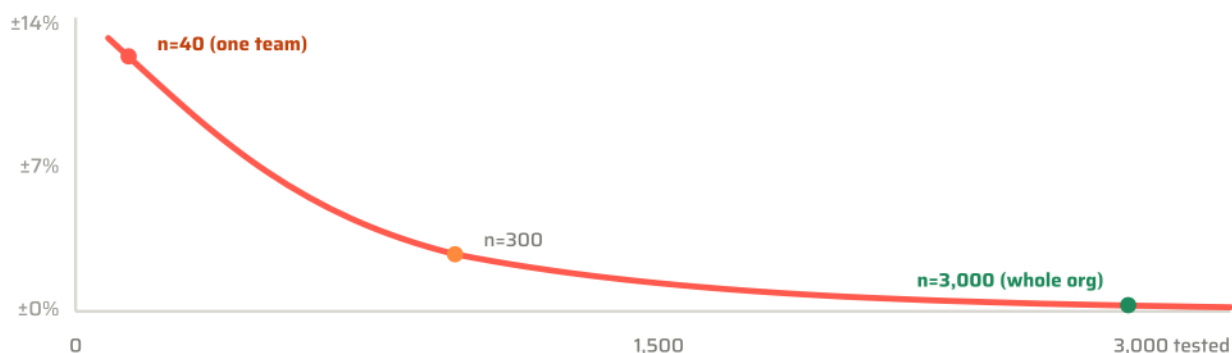
2.3 · SAMPLE SIZE & SEGMENTATION

SMALL TEAMS NEED MORE TESTS, NOT FEWER

Confidence in a measured rate depends on how many people were tested. This creates a trap for exactly the segmentation a CISO most needs. The organization-wide number may be stable, but the moment you slice it by a 40-person department — precisely where targeted action lives — the sample collapses and the uncertainty explodes. One annual test per small team is statistical guesswork.

Margin of error vs. number tested

95% confidence, rate near 20% · more tested → tighter estimate



The resolution is counterintuitive but decisive: **test small segments more often, not less**. Accumulating many small tests across a team over a quarter builds the sample size that a single annual test can never reach, making department-level and even role-level risk statistically legible. Continuous testing is the only way to measure the granular risk that targeted remediation depends on.

±12%

error on a 40-person team from one test — uselessly wide.

±3%

same team across a quarter of weekly micro-tests.

4×

tighter estimate from the same people, just tested more often.

2.4 · THE NEGATIVE-TRAINING TRAP

A BADLY RUN TEST CAN RAISE RISK

Beyond measuring poorly, the legacy approach can actively backfire. Punitive “gotcha” tests, public shaming of clickers, and unrealistic lures designed to maximize failure all corrode the one behavior that matters most: reporting. When employees fear punishment, they stop reporting suspicious messages — and the security team loses its earliest, cheapest signal of a live attack.

Suppressed reporting

Fear of blame teaches people to hide mistakes, blinding the SOC to real intrusions in progress.

Gamed results

Predictable, easy tests train people to spot the test — not the threat — inflating scores while real risk stands still.

Eroded trust

Adversarial testing frames security as the enemy, dissolving the culture a program depends on.

The fix

Blameless testing that rewards reporting, coaches privately, and treats a click as a data point — not a verdict.

A measurement instrument must not perturb the thing it measures. A test program designed to humiliate changes behavior in the wrong direction and contaminates its own data. The credible alternative is the opposite in spirit: realistic but fair, continuous but blameless, and oriented toward building reporting as the primary defensive behavior.

CHAPTER 2 · KEY TAKEAWAYS

- ◆ A single annual test is one noisy sample; its confidence interval hides most meaningful change.
- ◆ Whatever resilience a test builds decays within a quarter, leaving most of the year exposed.
- ◆ Department-level risk needs many small tests to reach a usable sample size.
- ◆ Punitive testing suppresses reporting and gamed easy tests inflate scores — both raise real risk.

03

ADVERSARY-REALISTIC SIMULATION

If the test does not look like the attack, it does not measure the risk. This chapter defines realism – across the kill chain and across every channel.

3.1 · THREAT EMULATION PRINCIPLES

TEST AGAINST THE ATTACKER, NOT A TEMPLATE

The validity of a phishing measurement rests entirely on how closely the test resembles a real attack. A lure that is obviously a drill — stale theme, clumsy pretext, generic greeting — measures only whether employees can spot a drill. To measure actual exposure, the simulation must emulate the tradecraft of a live adversary: current themes, plausible context, real-world pretexts, and the same psychological levers attackers pull.

Current themes

Lures track what attackers are using now — payroll changes, MFA resets, vendor invoices, executive requests.

Plausible context

References to real tools, teams, and processes the recipient would expect to see.

Psychological levers

Authority, urgency, fear, and curiosity — applied as a real social engineer would.

Evasive craft

Lookalike domains, hidden redirects, and QR payloads that mirror how attacks slip past controls.

The realism dial

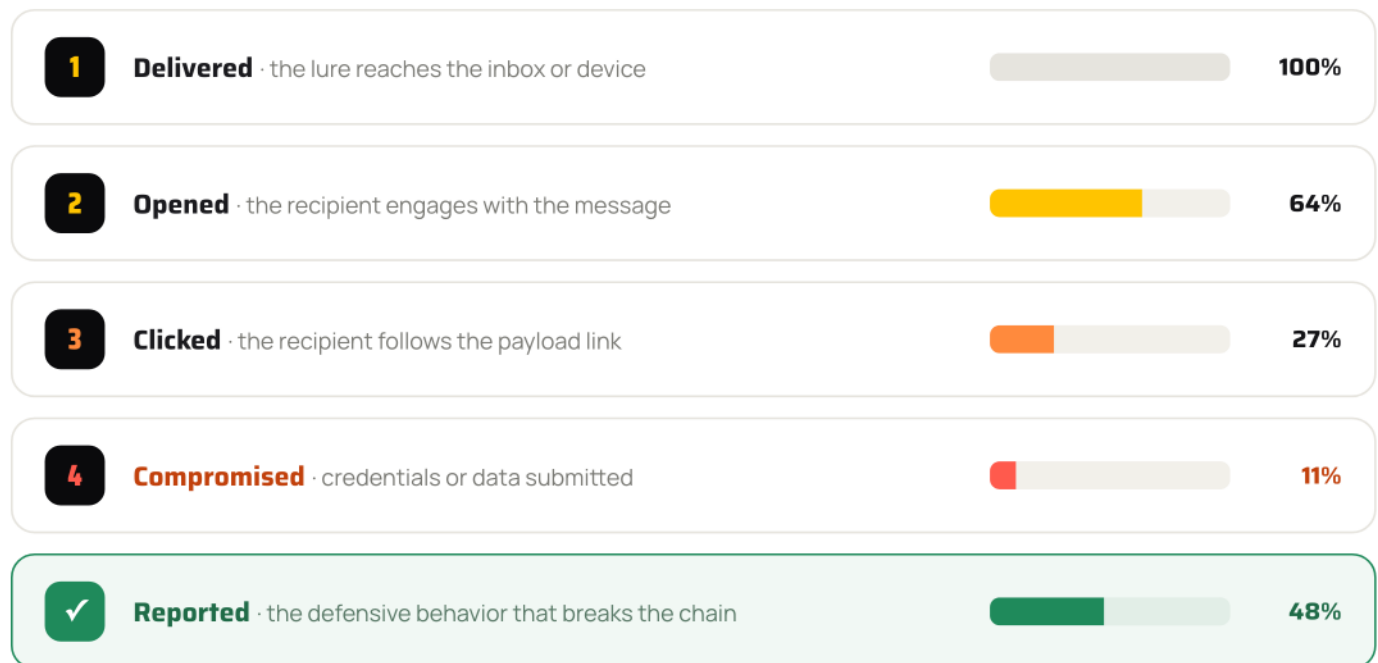
Realism is not a single setting but a spectrum the program must control deliberately. Too easy, and the test flatters the organization while teaching nothing. Too hard, too soon, and it demoralizes. The right design starts calibrated to the population's current resilience and escalates as that resilience grows — always staying a step behind the real attacker so practice remains both fair and meaningful.



3.2 · THE PHISHING KILL CHAIN

MEASURE EVERY STAGE, NOT JUST THE CLICK

A real phishing attack is a sequence, and a user can be stopped — or measured — at any stage. Recording only the click discards most of the signal. A rigorous program instruments the whole chain: who opened, who clicked, who submitted credentials, and crucially, who reported. Each transition is a distinct behavior with its own risk weight.



Instrumenting the full chain transforms the data. A team with a high click rate but an even higher report rate is far safer than its click rate alone suggests — and only a program that measures both can see it. **Reporting is the stage that matters most**, because a reported attack is a contained attack, regardless of how many people clicked.

3.3 · OMNICHANNEL TESTING

PROBE EVERY DOOR, NOT JUST THE FRONT ONE

Because attackers operate across email, SMS, voice, and QR, a measurement confined to email reports a flattering fiction. The fastest-growing vectors are precisely the ones employees are least prepared for, so a multi-channel program typically uncovers **higher** risk than email-only testing — not because the organization got worse, but because it is finally measuring what was always there.



Email phishing

Classic credential harvesting and BEC; the well-defended channel where most testing already lives.



Smishing (SMS)

Bypasses email gateways and lands on a personal device with no corporate filtering.



Vishing (voice)

Increasingly powered by cloned audio; tests resilience to live, real-time social pressure.



Quishing (QR)

Sails past link scanners as an image, then moves the victim onto an unmanaged phone.

The measurement principle: exposure can only be quantified for channels you actually test. Anything untested is unmeasured — and unmeasured risk is, by default, assumed to be zero, which is the most dangerous assumption a CISO can make.

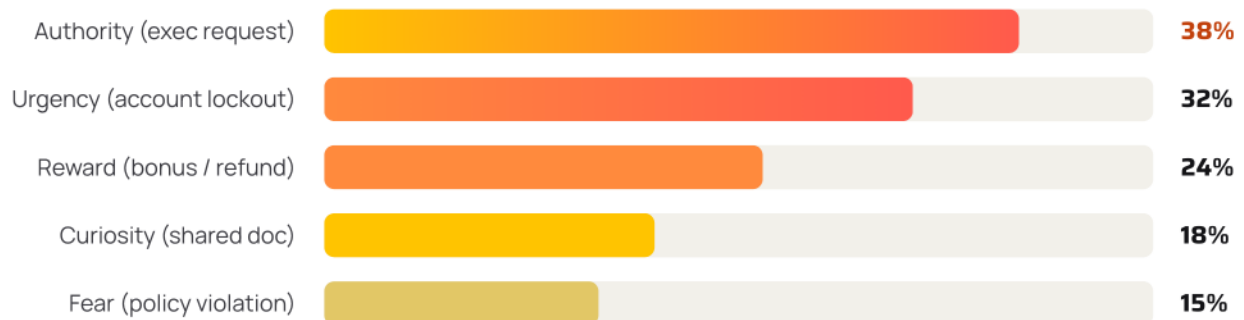
3.4 · THE LEVERS, MEASURED

WHICH MANIPULATIONS ACTUALLY WORK ON YOUR PEOPLE

A realistic program does more than measure whether people click – it measures **why**. By systematically varying the psychological lever behind each lure, the program reveals which manipulations are most effective against a given population, turning social-engineering theory into an organization-specific risk map.

Click rate by psychological lever

Share of recipients who acted, by lure type · illustrative population



This intelligence directs the program. If authority-based lures dominate, executive-impersonation defenses and out-of-band verification habits become the priority; if reward lures land, finance and procurement need focused work.

The lever map turns a generic curriculum into a targeted intervention, spending effort exactly where the population is most manipulable.

3.5 · WHAT REALISM BUYS YOU

VALIDITY IS THE WHOLE POINT

Every design choice in this chapter serves a single goal: a measurement valid enough to bet a budget on. A test that looks like an attack, spans every channel, instruments the full kill chain, and varies its psychological levers produces data that genuinely predicts real-world exposure. A test that does none of these produces a comforting number with no predictive power.

Valid

Realistic tests predict real behavior under real attack.

Complete

Every channel and stage is instrumented, not just the click.

Diagnostic

The lever map shows exactly where to intervene.

Realism, however, raises a hard operational question. Producing fresh, current, multi-channel, lever-varied lures at scale — and keeping pace with attackers who now use AI — is far beyond what a human team can sustain by hand. Meeting the standard this chapter sets requires an engine purpose-built for it. That engine is the subject of Chapter 4.

CHAPTER 3 · KEY TAKEAWAYS

- ◆ A test is only valid if it emulates real attacker tradecraft; obvious drills measure nothing useful.
- ◆ Instrument the whole kill chain — reporting is the stage that breaks the attack.
- ◆ Test every channel; unmeasured channels are silently assumed risk-free.
- ◆ Varying the psychological lever turns a generic program into a targeted one.

04

THE CYBEARS X PHISH ENGINE

Realism at scale is an engineering problem. Here is the closed-loop architecture that generates adversary-grade attacks, targets them intelligently, and turns every result into the next test.

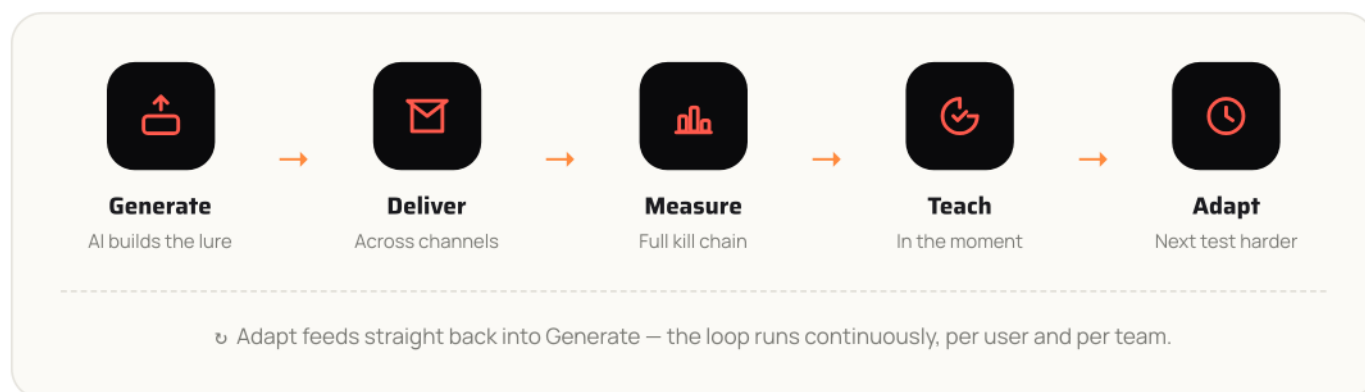
◆ Closed loop ◆ AI generation ◆ Targeting ◆ Automation



4.1 · CLOSED-LOOP ARCHITECTURE

TEST FEEDS TRAINING FEEDS TEST

X Phish is built as a closed loop in which every stage informs the next. A test produces behavioral data; that data updates each user's risk profile; the profile drives a teachable moment and the design of the next, harder test. There is no start and no end — only a continuously self-tuning instrument that gets sharper with every cycle.



Two consequences follow from the architecture. First, the program never goes stale: because generation is AI-driven, no two cycles repeat, defeating the "spot the test" gaming that plagues template libraries. Second, difficulty self-calibrates: users who keep reporting graduate to harder tests, while those who struggle receive more support — all without manual curation.

4.2 · AI ATTACK GENERATION

FIGHT AI ATTACKS WITH AI TESTS

Because real attackers now generate lures with AI, only an AI-driven test engine can keep pace. X Phish composes fresh, current, context-aware lures on demand — varying theme, channel, pretext, and psychological lever — so the simulation reflects the sophistication of live threats rather than a frozen library written months ago. The engine operates within strict guardrails: realistic but safe, never exposing real credentials or sensitive data.

GENERATED PER LURE

- Subject line & body copy
- Sender identity & pretext
- Brand styling & landing page
- Language & localization
- Difficulty calibration

GUARDRAILS

- No real credential capture
- No sensitive-data exposure
- Tone & content policy filters
- Approval workflow & audit log
- Region & residency controls

The result is an effectively infinite, always-current library that no employee can memorize and no attacker trend can outdate. Crucially, generation is conditioned on each recipient's profile and the organization's own context, so the test a finance manager receives differs from the one a developer receives — both plausible, both calibrated, both measuring the risk that actually applies to that person.

4.3 · TARGETING & ADAPTIVE DIFFICULTY

PRESSURE WHERE THE RISK IS

Uniform testing wastes effort on the resilient and under-tests the vulnerable. X Phish concentrates pressure dynamically: each user's demonstrated behavior sets the difficulty and frequency of their next test, so the system spends its energy precisely where risk concentrates — the small minority who drive most of the exposure identified in Chapter 1.

HIGH-RISK USER

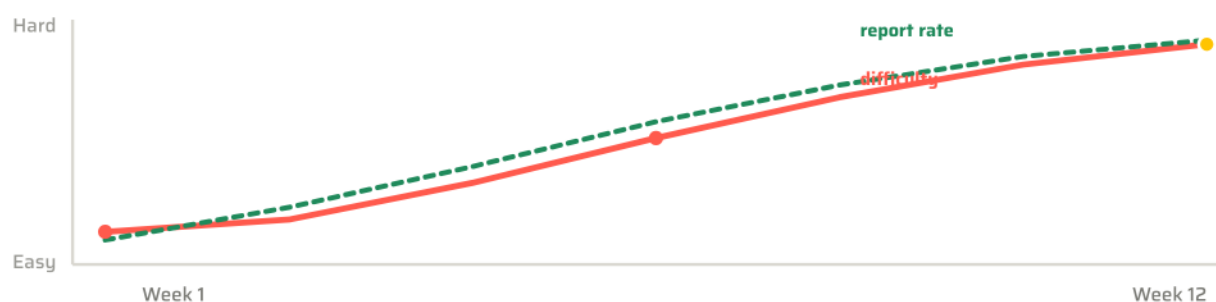
Repeat clicker in finance → higher test frequency, escalating realism, targeted remediation, and a tighter watch on the report rate until behavior turns.

LOW-RISK USER

Consistent reporter → occasional APT-grade tests to keep skills sharp, minimal interruption otherwise. Resilience is confirmed, not assumed.

How a high-risk user's difficulty adapts over time

Test difficulty rises only as the user's report rate climbs · illustrative

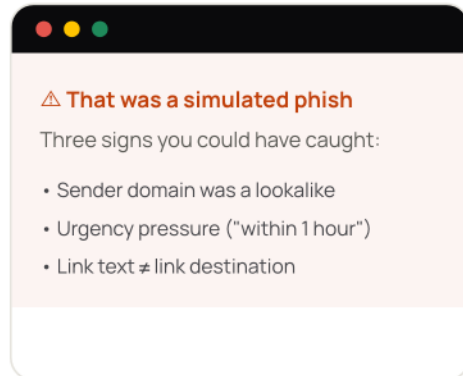


This adaptivity is what makes a continuous program humane as well as effective. Resilient employees are not nagged; vulnerable ones are not thrown straight into APT-grade lures they cannot yet handle. The difficulty curve always tracks the individual, keeping every test fair, informative, and — because it never plateaus below real-attack sophistication — genuinely protective.

4.4 · THE TEACHABLE MOMENT & AUTOMATION

THE CLICK IS THE LESSON

When a user interacts with a simulated lure, X Phish does not log a failure and move on. It delivers an immediate, blameless micro-lesson at the peak of attention — pointing out the exact signals that gave the lure away. This is the single most memorable lesson the user will receive, and it closes the loop between measurement and improvement in real time.



- ◆ Auto-enrolment & directory sync — joiners, movers, and leavers update the test population automatically.
- ◆ Risk-triggered remediation — a click assigns the right follow-up lesson with no human in the loop.
- ◆ Campaign orchestration — continuous, spaced, multi-channel scheduling that runs itself.
- ◆ Live reporting — dashboards and board exports refresh in real time as results land.

Automation is what makes the standard sustainable. A single administrator can operate a sophisticated, continuous, omnichannel program across thousands of employees, because enrolment, generation, delivery, teaching, and reporting all run end-to-end. The lean security team gets enterprise-scale rigor without an enterprise-scale headcount.

CHAPTER 4 · KEY TAKEAWAYS

- ◆ A closed loop keeps the program current and self-calibrating — no two cycles repeat.
- ◆ AI generation matches attacker speed within strict safety guardrails.
- ◆ Adaptive targeting concentrates effort on the users who drive most of the risk.
- ◆ Real-time teaching and full automation make enterprise scale a one-person job.

4.5 · THE ENGINE AT A GLANCE

CAPABILITY SUMMARY

Every property the preceding chapters argued for is realized as a concrete capability. The table below maps the requirement to the engine feature that satisfies it — the bridge from principle to product.

REQUIREMENT	X PHISH CAPABILITY
Continuous	Always-on campaign orchestration with spaced scheduling.
Adversary-realistic	AI-generated, current, lever-varied lures within safety guardrails.
Omnichannel	Email, SMS, voice, and QR simulation from one console.
Statistically sound	High-frequency micro-tests build sample size for segment-level confidence.
Financially expressed	Behavioral data feeds the dollar-denominated Human Risk Score (Chapter 5).
Closed-loop	Test → teach → adapt → retest, fully automated and self-tuning.

4

channels simulated

5

kill-chain stages tracked

∞

unique AI lures

1

admin to run it all

05

QUANTIFYING RISK IN DOLLARS

The payoff of rigorous measurement: a single, trended, dollar-denominated number the board reads as exposure and the CISO reads as progress.

5.1 · THE HUMAN RISK SCORE

ONE COMPOSITE, FOUR SIGNALS

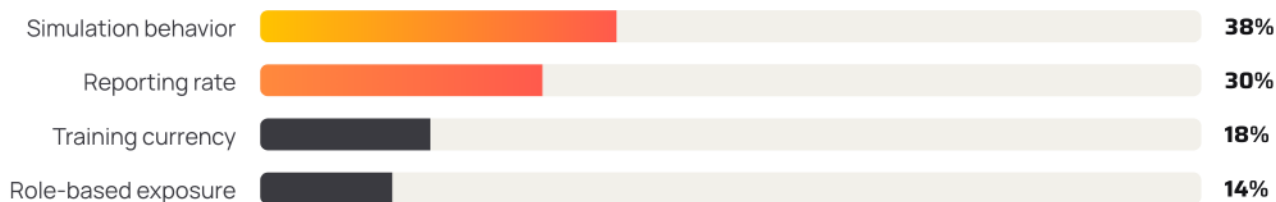
The Human Risk Score (HRS) condenses the kill-chain data of Chapter 3 into one trended measure of human exposure. It weights behavior under realistic simulation most heavily, because behavior under pressure is the best predictor of real-world compromise — while still crediting reporting, training currency, and role-based exposure.

B+

SCORE 78 / 100

What the grade represents

A weighted composite of simulation behavior, real-world reporting, training currency, and role exposure — a letter grade for the board and a precise score for the analyst, both mapped to a dollar figure.

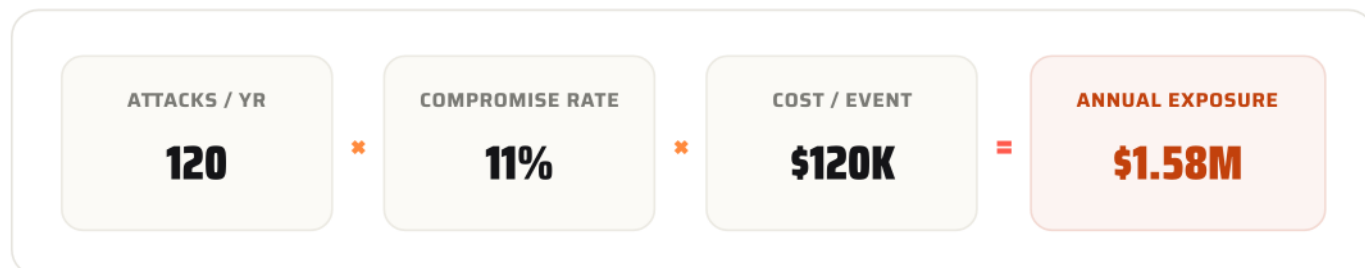
Inputs & weighting

Because the score is computed from continuous, high-frequency tests, it carries the statistical confidence Chapter 2 showed a single annual test cannot. The HRS is not a vanity metric; it is a measured estimate with a known margin of error — which is precisely what makes the dollar translation that follows defensible.

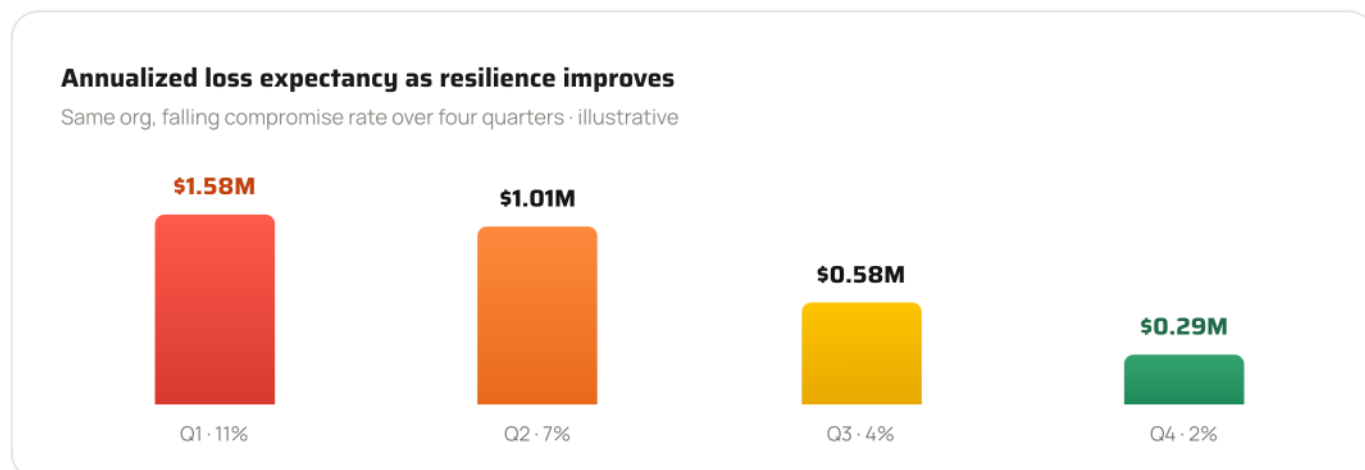
5.2 · FROM CLICK RATE TO LOSS EXPOSURE

THE RISK EQUATION, MADE EXPLICIT

Translating behavior into money follows a transparent chain. The measured compromise rate sets the probability that a real attack succeeds; multiplied by attack frequency and the expected cost of a successful event, it yields an **annualized loss expectancy (ALE)** — the same model insurers and risk teams already use. Every term is observed, not guessed: X Phish supplies the compromise and report rates directly.



How reducing the click rate moves the money



Reframing the click rate as ALE changes the conversation entirely. A four-point drop in compromise is no longer an abstract percentage; it is **over a million dollars of avoided expected loss** — a number that competes for budget on equal footing with every other investment the board weighs.

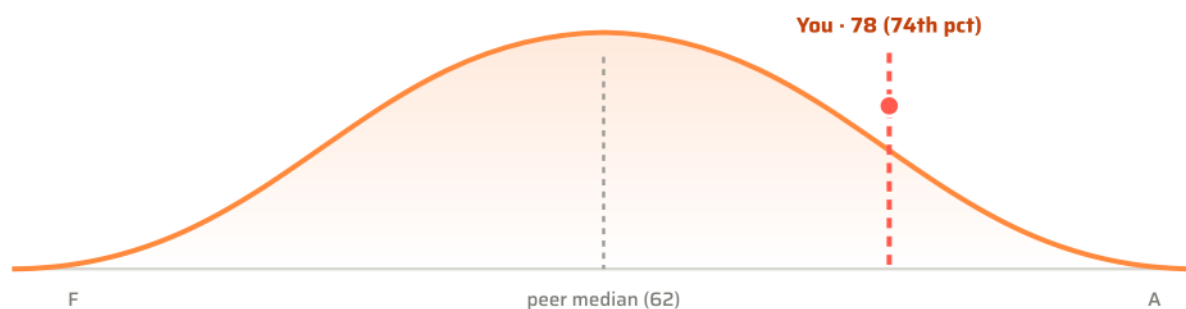
5.3 · BENCHMARKING

WHERE YOU STAND AGAINST YOUR PEERS

A score gains meaning from context. X Phish places an organization's HRS within the distribution of comparable organizations — by sector, size, and region — so leadership can answer not only "are we improving?" but "are we ahead or behind those we are compared to?" Percentile position is often the figure a board remembers most.

Human Risk Score distribution across peer organizations

Your position vs. the peer curve · illustrative



VS. SELF

Quarter-over-quarter trend proves the program works.

VS. TEAMS

Surfaces the departments that need attention.

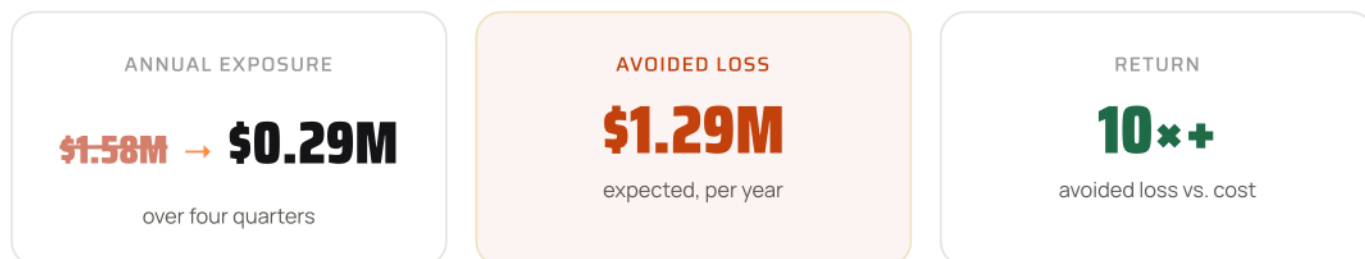
VS. PEERS

Positions the org against its sector for the board.

5.4 · BOARD REPORTING & ROI

A PROGRAM THAT PAYS FOR ITSELF, PROVABLY

With exposure expressed in dollars, the return on the program is a subtraction, not a leap of faith. The avoided loss from a falling compromise rate is set against the cost of the platform — and for any organization of meaningful size, the avoided loss dominates by an order of magnitude. The CISO walks into the board meeting with the one slide that always lands: risk down, dollars saved, cost justified.



This is the strategic destination of the entire whitepaper. Rigorous, continuous, realistic measurement is not an academic exercise — it is what lets human risk be governed like any other enterprise risk: quantified, trended, benchmarked, priced, and managed down with a demonstrable return.

CHAPTER 5 · KEY TAKEAWAYS

- ◆ The HRS condenses kill-chain data into one trended, statistically-grounded measure.
- ◆ The ALE model turns the click rate into a dollar figure from observed, not guessed, inputs.
- ◆ Peer benchmarking gives the board the percentile position it remembers.
- ◆ Avoided loss vs. cost makes ROI a subtraction — typically 10x or better.

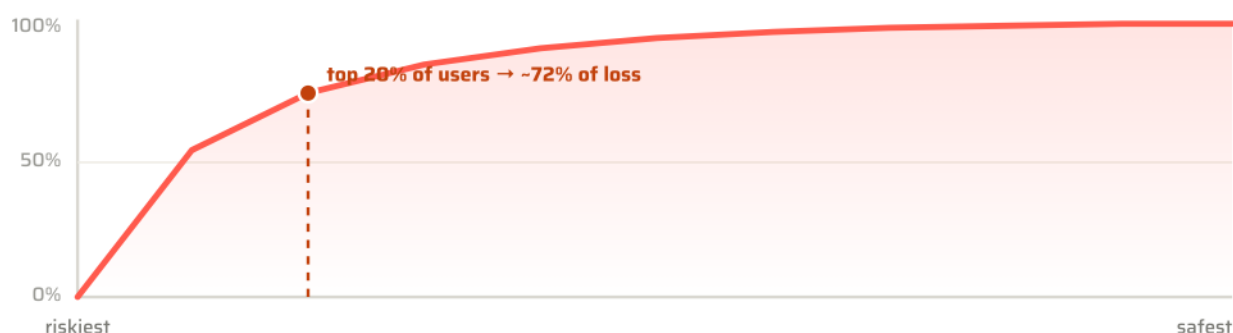
5.5 · WHERE THE EXPOSURE CONCENTRATES

A SMALL MINORITY, MOST OF THE RISK

Quantification reveals a structure that uniform programs miss entirely: human risk is heavily concentrated. A minority of repeat-clicking users accounts for the large majority of expected loss. This Pareto shape is the single most actionable finding a measured program produces, because it tells leadership exactly where a dollar of effort buys the most risk reduction.

Cumulative share of expected loss by user risk decile

Highest-risk users on the left · illustrative



The operational implication is profound. Rather than spreading effort evenly — over-training the resilient and under-protecting the vulnerable — a measured program directs its adaptive pressure at the high-loss minority. **Targeting the top two deciles can retire the majority of an organization's human-risk exposure** at a fraction of the cost of a uniform campaign, which is exactly what the X Phish targeting engine is built to do.

06

EVIDENCE & OUTCOMES

The model, observed. Longitudinal results, a controlled cohort comparison, a department risk map, and a representative deployment.

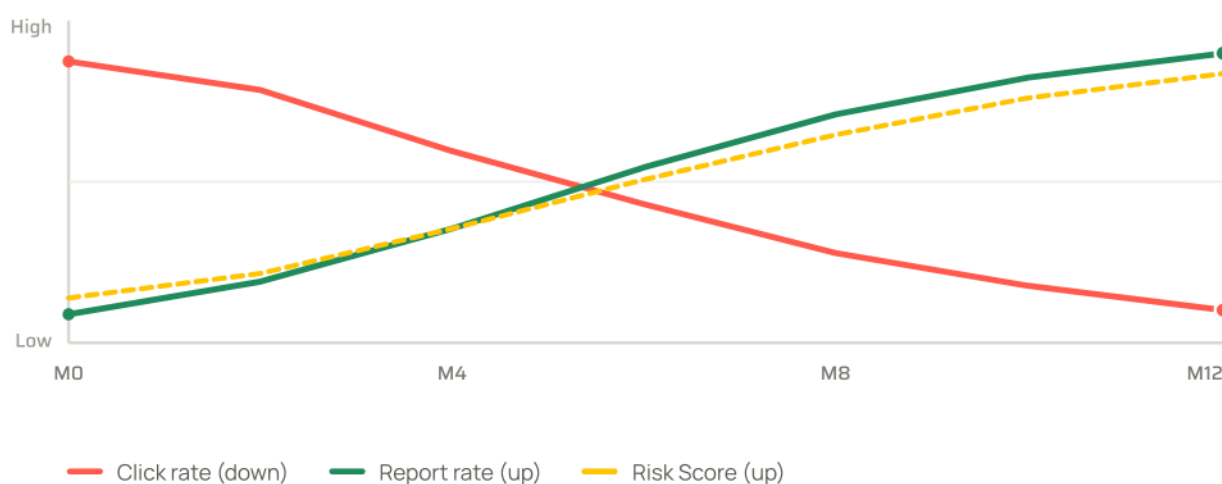
6.1 · LONGITUDINAL RESULTS

TWELVE MONTHS, THREE MOVING LINES

The clearest evidence of a working program is the simultaneous movement of three metrics over time: the click rate falls, the report rate rises, and the two together pull the Human Risk Score upward. Plotted on one timeline, they tell the whole story — exposure shrinking as the workforce turns from liability into sensor.

Click rate, report rate & risk score over 12 months

Program launched in month 0 · illustrative deployment



29% → 5%

click rate

7% → 64%

report rate

D → A-

risk score

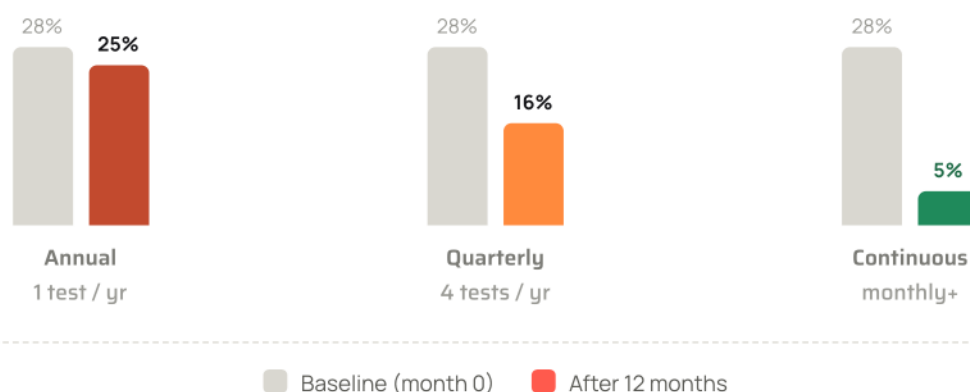
6.2 · A CONTROLLED COHORT COMPARISON

FREQUENCY IS THE VARIABLE THAT MATTERS

To isolate the effect of cadence, consider three matched cohorts drawn from the same organization, differing only in how often they are tested. The contrast is stark and consistent with the statistics of Chapter 2: the annual cohort barely moves, the quarterly cohort improves, and the continuous cohort transforms.

Compromise rate after 12 months, by testing cadence

Matched cohorts, same population · illustrative controlled comparison



The annual cohort's three-point improvement sits well within the statistical noise of Chapter 2 — it may be no improvement at all. The continuous cohort's collapse from 28% to 5% is large, sustained, and far outside the margin of error. **Cadence, not content, is the dominant lever**, and only continuous testing delivers a result a CISO can defend to the board.

6.3 · THE RISK MAP

SEE EXACTLY WHERE TO ACT

Aggregated numbers hide the structure leadership needs. A department-by-lever heatmap exposes it: each cell is the click rate for one team against one manipulation type, so the hotspots — and the safe zones — are visible at a glance. This is the operational artifact that turns measurement into a remediation plan.

Click rate by department & lure type

	Authority	Urgency	Reward	Curiosity
Finance	41%	34%	22%	14%
Executive	33%	24%	12%	9%
Sales	26%	23%	31%	17%
Engineering	11%	13%	8%	21%
Operations	18%	25%	15%	10%

Lower risk  Higher risk

The map writes the action plan automatically. Finance under authority lures is the brightest hotspot — so executive-impersonation defenses there become priority one. Sales is uniquely exposed to reward lures; engineering to curiosity. **Each cell points to a specific, targeted intervention** rather than a blanket campaign, concentrating effort exactly where the measured risk lives.

6.4 · A REPRESENTATIVE DEPLOYMENT

A 4,000-PERSON ENTERPRISE, TWELVE MONTHS

The following vignette illustrates the pattern this whitepaper describes — a regional enterprise of roughly 4,000 employees moving from a single annual test to a continuous, multi-channel X Phish program. It is representative of the trajectory continuous programs report, not a guarantee; outcomes depend on starting point, sponsorship, and consistency.

- Month 0 · Baseline**
Multi-channel baseline reveals a 29% click rate, a 7% report rate, and an HRS of D — well below peer median, with finance and sales as the worst hotspots.
- Months 1-3 · Target the hotspots**
Adaptive targeting concentrates on the high-risk minority; teachable moments fire on every click. Click rate falls into the mid-teens.
- Months 4-9 · Build the sensor**
One-click reporting and positive reinforcement push the report rate past 50%; real attacks now surface in minutes.
- Month 12 · Quantified outcome**
Click rate 5%, report rate 64%, HRS A- (74th percentile). Annualized loss exposure cut from \$1.58M to under \$0.30M.

"The number that changed the board's mind wasn't the click rate. It was seeing \$1.3 million of expected loss come off the table — measured, not asserted."

6.5 · WHAT THE EVIDENCE SHOWS

THE MODEL HOLDS UP

Across the longitudinal trend, the controlled cohort comparison, the risk map, and the representative deployment, one conclusion recurs: a continuous, realistic, quantified program reduces measured human risk by a margin that is large, sustained, and statistically defensible — and it expresses that reduction in dollars the board can act on.

-83%

relative reduction in click rate over twelve months of continuous testing.

9×

increase in the report rate — the workforce becomes a sensor.

72%

of loss exposure traced to the top 20% of users — target them first.

\$1.3M

representative annual avoided loss, measured as ALE reduction.

CHAPTER 6 · KEY TAKEAWAYS

- ◆ Click down, report up, score up — three lines that together tell the whole story.
- ◆ A controlled cohort comparison shows cadence is the dominant lever, not content.
- ◆ The department×lever heatmap converts measurement into a targeted remediation plan.
- ◆ A representative deployment converts an 83% click-rate drop into ~\$1.3M of avoided loss.

07

DEPLOYMENT, SOVEREIGNTY & COMPLIANCE

A measurement instrument must fit your governance — where data lives, which frameworks it evidences, and how it scales across many tenants.

7.1 · DEPLOYMENT & RESIDENCY

SOVEREIGN AND MEA-NATIVE BY DESIGN

Phishing simulation touches sensitive territory — employee behavior, internal context, and message content — so where that data lives is a first-order decision, not an afterthought. X Phish ships in three deployment models with identical capability, differing only in data location and operation, and is built MEA-native so regional residency and language requirements are met from day one rather than retrofitted.

SaaS Cloud

FASTEST TO LAUNCH

Fully managed and continuously updated. Live in days, no infrastructure to maintain — right for most organizations.

On-Premise

FULL CONTROL

Runs entirely inside your data center, behind your controls — for organizations whose policy keeps all systems in-house.

Sovereign-Hosted

IN-COUNTRY DATA

Hosted in-region by a sovereign partner to satisfy national residency mandates — without losing the managed experience.

In-region

data residency & hosting

Arabic

first-class localization

Day 1

sovereign readiness

7.2 · COMPLIANCE MAPPING

EVIDENCE OF EFFECTIVENESS, NOT JUST ACTIVITY

Phishing testing is an explicit requirement across the major frameworks, and auditors increasingly want proof that the control *works*, not merely that it exists. Because X Phish records every interaction continuously, the evidence an assessor asks for is a by-product of normal operation — exportable in a click, framework-mapped, and timestamped.

FRAMEWORK	CONTROL	EVIDENCE PRODUCED
NIST CSF 2.0	PR.AT — Awareness & Training	Continuous simulation records with behavioral trend.
ISO/IEC 27001	A.6.3 — Awareness & training	Per-employee participation and effectiveness data.
PCI DSS 4.0	Req. 12.6 — Security awareness	Phishing-specific testing with retained evidence.
GDPR	Art. 32 — Security of processing	Demonstrable staff resilience as an org measure.
DORA / NIS2	ICT risk & testing	Continuous human-risk testing with audit trail.
SOC 2	CC1 / CC2 — Control environment	Functioning awareness control with ongoing comms.

The dual record — that the control is both designed and effective — is exactly what modern audits demand, and it is the natural output of continuous, behavior-based measurement. Compliance stops being a periodic scramble and becomes an automatic export.

7.3 · MSSP & MULTI-TENANCY

ONE CONSOLE, MANY ORGANIZATIONS

For managed security providers and group functions overseeing many entities, X Phish runs as a multi-tenant platform: isolated tenants under one operational pane, with per-tenant branding, residency, and reporting. The same automation that lets one admin run one enterprise lets one team run dozens — each with its own quantified, benchmarked Human Risk Score.

Tenant isolation

Hard data separation per client, with independent residency and access control.

Cross-tenant view

A portfolio dashboard ranks every client's HRS for the provider's leadership.

White-label

Per-tenant branding on portals, reports, and teachable moments.

Scales linearly

Automation means adding a tenant adds almost no operational load.

CHAPTER 7 · KEY TAKEAWAYS

- ◆ Three deployment models cover SaaS, on-prem, and sovereign-hosted residency.
- ◆ MEA-native residency and Arabic localization are built in, not bolted on.
- ◆ Continuous records make framework evidence an automatic, one-click export.
- ◆ Multi-tenancy lets one team run dozens of quantified programs at once.

08

CONCLUSION & REFERENCES

Where the argument lands, how to begin, and the evidence base behind it.

8.1 · CONCLUSION

MEASURE THE HUMAN ATTACK SURFACE — THEN MANAGE IT DOWN

The argument of this whitepaper is a single, connected chain. Phishing is the dominant, AI-accelerated threat to the enterprise. The standard response — an annual test and a bare click rate — fails both as training and, more fundamentally, as measurement: it is one noisy sample of a moving target. A credible alternative exists: adversary-realistic, continuous, omnichannel, statistically sound testing that instruments the full kill chain.

From that foundation comes the decisive output. Behavioral data, measured with real statistical confidence, can be translated into a **dollar-denominated, trended, benchmarked estimate of human risk**. The click rate becomes annualized loss expectancy; the loss expectancy falls quarter over quarter; and the reduction — large, sustained, and defensible — is expressed in the only unit the board ultimately weighs. Human risk finally joins every other enterprise risk: quantified, governed, and managed down with a demonstrable return.

Cybears X Phish exists to operationalize that chain end to end — AI-generated omnichannel attacks, adaptive targeting, real-time teaching, and a Human Risk Score the board reads as exposure. The technology is the enabler; the outcome is a workforce transformed from the organization's largest attack surface into its most reliable sensor.

Realistic

Tests that mirror the real attack.

Continuous

Statistically valid, always-on.

Quantified

Risk priced in dollars.

START WITH A BASELINE

The first step is the most revealing: run a multi-channel baseline against your organization and receive your starting Human Risk Score — and the dollar exposure behind it. From there, every improvement is measured.

GET IN TOUCH

[\[email protected\]](#)

xphish.cybears.net

REFERENCES & METHODOLOGY

SOURCES & NOTES

Statistics in this paper are drawn from widely published industry research on phishing volume, breach costs, and human factors, together with established measurement and behavioral theory. Figures are presented as indicative ranges consistent with that body of work; charts are illustrative renderings of typical distributions rather than data from a single named source. Readers should consult the primary sources for precise, current numbers in their sector and region.

1. Verizon, *Data Breach Investigations Report* — annual analysis of breach patterns and the human element in initial access.
2. FBI Internet Crime Complaint Center (IC3), *Internet Crime Report* — reported business email compromise losses by year.
3. IBM Security, *Cost of a Breach Report* — breach cost components, containment time, and the effect of preparedness.
4. Anti-Phishing Working Group (APWG), *Phishing Activity Trends Report* — global phishing volume and channel evolution.
5. R. Cialdini, *Influence: The Psychology of Persuasion* — the levers (authority, urgency, scarcity, social proof) social engineering exploits.
6. H. Ebbinghaus, *Memory: A Contribution to Experimental Psychology* — the forgetting curve and the spacing effect.
7. D. Kahneman, *Thinking, Fast and Slow* — dual-process cognition and the heuristics attackers target under load.
8. Open FAIR™ (The Open Group) — the quantitative risk model underlying annualized loss expectancy (frequency × magnitude).
9. NIST, *Cybersecurity Framework 2.0*; ISO/IEC 27001:2022; PCI DSS 4.0; EU DORA & NIS2 — control requirements for awareness, testing, and ICT risk.
10. ENISA & regional cyber agencies — guidance on phishing-resilience programs and AI-enabled social engineering.

A note on figures

The longitudinal results, cohort comparison, risk map, and deployment vignette are representative of the patterns organizations report when adopting continuous, behavior-based phishing programs; they illustrate the model rather than promise a specific outcome. Dollar figures use the Open FAIR / ALE structure (attack frequency × compromise rate × cost per event) with indicative inputs; substitute your own measured values for an organization-specific estimate.



SEE YOUR HUMAN RISK, IN **DOLLARS**

Run a multi-channel baseline and receive a board-ready Human Risk Score — and the loss exposure behind it.

Request a baseline assessment

xphish.cybears.net

[\[email protected\]](#) · Your Cyber Resiliency Partner · Hydra, Algiers