

DOI: https://doi.org/10.48009/2_iis_2026_147

Steady-state residual loss in socio-technical phishing defense

Joshua Taylor, *Dakota State University*, Josh.Taylor@trojans.dsu.edu

Dr. Omar El-Gayar, *Dakota State University*, Omar.El-Gayar@dsu.edu

Dr. Abeer Alkhwalidi, *Dakota State University*, Abeer.Alkhwalidi@dsu.edu

Abstract

Phishing remains a leading initial-access vector despite sustained investment in user training and technical filtering. The literature on training decay, bounded technical controls, and adversarial adaptation each treat one slice of the problem in isolation. We develop a discrete-time simulation that couples all three subsystems within a unified socio-technical framing, deliberately modeling the adversary as a minimally adaptive multi-armed bandit. The choice is methodological: a minimal-adaptation adversary establishes a conservative lower bound on adversarial capability, so that more sophisticated attackers, including those driven by large language models, can be expected only to elevate the resulting steady-state loss floor. Across 200 random seeds and a 21-point control-strength sweep, the floor is structurally positive in every condition tested; the adaptive premium over a uniform-random baseline is approximately 12% and persists across alternative learning algorithms, feedback structures, and an enlarged action space; and the loss-versus-control-strength relationship is approximately linear in residual exposure. The bandit's selections concentrate on high-quality, high-personalization actions, the operational mechanism through which AI-enabled adversaries elevate the floor in our framework. We interpret the floor as the planning baseline against which the agentic-AI adversarial landscape should be evaluated, not as a ceiling.

Keywords: phishing defense, socio-technical systems, multi-armed bandit, adversarial adaptation, AI-enabled threats, security simulation

Introduction

Phishing remains a predominant initial-access vector in organizational cyber incidents, accounting for a substantial share of reported breaches across recent industry analyses (Verizon, 2024; IBM Security, 2024). Despite sustained investment in security awareness training and probabilistic technical controls, residual phishing loss persists at non-trivial rates across operational reports. Three streams of research separately address why. The first stream, anchored in the cognitive psychology of forgetting, documents that training-induced reductions in susceptibility decay over time absent reinforcement (Ebbinghaus, 1885; Murre and Dros, 2015; Canfield et al., 2016). The second stream, anchored in security-control engineering, documents that technical filters are probabilistic rather than deterministic, with detection rates that vary by campaign novelty and targeting (Verizon, 2024; Heiding et al., 2024). The third stream, anchored in adversarial reasoning and game theory, documents that attackers adapt strategy in response to defender posture (Tambe, 2011; Laszka et al., 2014; Nguyen and Reddi, 2023). Each stream is empirically grounded; each stream answers a partial question.

What is missing from the existing literature is a tractable account of what happens when these three mechanisms operate jointly. The training-decay literature does not model how attacker selection responds to user susceptibility windows. The technical-controls literature does not model how repeated exposure shapes the probability that any given attempt becomes a clicked-and-uncontrolled incident. The adversarial-adaptation literature treats the defender side as a static target rather than as a dynamic socio-technical system. Without an integrated model, the practitioner question “is residual phishing loss a structural feature of joint defender-attacker dynamics, or is it an artifact of insufficiently aggressive component-level investment?” cannot be answered.

We address this gap by developing a discrete-time simulation that couples a logistic susceptibility model with parameterized training decay, a probabilistic technical mitigation layer with bounded effectiveness, and a minimally adaptive adversary modeled as an epsilon-greedy multi-armed bandit. The simulation is parameterized to literature-anchored ranges where empirical anchors exist, and is honest about stylization where they do not. The minimal adversary serves as a conservative lower bound on adaptive capability: if persistent loss emerges under epsilon-greedy adaptation with feedback only on realized loss, then more sophisticated adversarial reasoning, including reinforcement-learning-driven content generation and large-language-model-assisted personalization, can only elevate that loss.

Research Question. Under jointly modeled training decay, bounded technical controls, and minimal adversarial adaptation, what is the steady-state residual loss floor, and how robustly does it persist across modeling choices?

The paper makes three contributions. The first is methodological. We introduce a conservative-lower-bound framing of the adversary subsystem in integrated phishing-defense models: by modeling the attacker as a minimally adaptive multi-armed bandit with epsilon-greedy selection over a discrete action space and reward feedback restricted to realized loss, we establish a floor on residual loss against which more sophisticated adversaries can be evaluated. The framing converts what is conventionally a limitation of simulation-based adversary modeling, namely that any specific learning algorithm is one of many possible choices, into a feature: if the floor is robust to the choice of algorithm, the floor is a property of the joint defender-attacker dynamics rather than an artifact of the algorithm. The second contribution is empirical. We demonstrate that the floor is robust across alternative learning algorithms (epsilon-greedy, Thompson sampling, softmax), feedback structures (realized loss versus click occurrence), and action-space sizes ($K = 9$ versus $K = 25$). The third contribution is interpretive. We map AI-enabled phishing onto the model as parameter movements rather than as a categorical break, identifying message quality, personalization, action-space size, and effective learning rate as the four channels through which large-language-model-driven adversaries elevate the floor.

Related Work

The empirical literature on phishing-awareness training establishes two robust facts. First, training reduces simulated-campaign click rates by a measurable margin: Sommestad and Karlzén (2019), in a meta-analysis of 48 phishing field experiments, report a central tendency of training effects in the 5 to 15 percentage-point range, consistent with earlier studies (Caputo et al., 2014; Jensen et al., 2017). Second, the gains decay with time. Canfield et al. (2016) document substantial decay in detection performance within four to six weeks post-training, and Murre and Dros (2015), replicating Ebbinghaus (1885), recover a power-function forgetting curve under which retention falls to roughly half within two to three weeks at workday timescales. Training is effective but not durable: organizational training cadences that match decay timescales are required to maintain operational benefit.

Modern phishing-control infrastructure layers email filtering, link rewriting, attachment sandboxing, and authentication challenges. The Verizon (2024) Data Breach Investigations Report documents that despite aggregate filtering performance well above ninety percent on routine traffic, novel and targeted campaigns evade detection at substantially higher rates, and that human-element-involved breaches remain a dominant initial-access pattern. Biggio and Roli (2018) frame the underlying mechanism as adversarial concept drift: as attackers continuously modify content to evade learned classifiers, classification remains probabilistic rather than deterministic. Heiding et al. (2024) provide a salient recent demonstration that large-language-model-generated phishing emails achieve click-through rates comparable to expert-crafted emails, with human recipients unable to reliably distinguish AI-generated from human-crafted content; the assumption of a sharp message-quality gradient between attacker and defender is no longer durable.

Game-theoretic security models, Stackelberg formulations in particular, treat attacker-defender interaction as sequential commitment under bounded information (Tambe, 2011; Laszka et al., 2014). Recent work relaxes the rationality assumption by modeling learning adversaries through reinforcement learning and online optimization (Nguyen and Reddi, 2023), establishing that adaptive adversaries can identify effective strategies through trial-and-error feedback even without complete environmental knowledge. The economic-security literature (Gordon and Loeb, 2002; Grossklags and Johnson, 2009) frames the implication as an investment problem: under adaptive pressure, optimal defensive spend yields diminishing marginal returns and non-zero residual risk.

Several prior agent-based simulation efforts engage organizational security or phishing dynamics. Burns et al. (2017) develop an ABM of organizational information-security behavior that examines how peer effects and management posture shape compliance. Abbasi et al. (2021) study a phishing funnel in which message characteristics, user attention, and detection technologies jointly determine click and report outcomes. We extend that tradition by introducing the conservative-lower-bound framing of the adversary subsystem and pairing it with a joint analysis of robustness across learning algorithms and feedback structures. The combination is what licenses an interpretation of residual loss as a structural floor against which more sophisticated adversaries can be evaluated, rather than as a point estimate dependent on the specific learning algorithm chosen.

We anchor the paper in the security-ergonomics tradition of socio-technical systems theory. Reason (2000) develops the Swiss cheese model: defensive layers each have individual weaknesses, and failure occurs when the gaps in successive layers align. Sasse and Flechais (2005) develop the usable-security perspective in which technical controls operate within social and behavioral contexts. Beautement, Sasse, and Wonham (2008) develop the compliance-budget framing in which users have a finite tolerance for security friction. These three threads, taken together, offer a richer foundation for thinking about jointly modeled defense than the older organizational-MIS framing of Bostrom and Heinen (1977). The empirical NIST Phish Scale work of Steves, Greene, and Theofanos (2020) provides operational anchors for thinking about how message difficulty and target characteristics jointly determine click probability, complementing the Somwestad and Karlzén (2019) meta-analytic synthesis.

The gap our paper addresses is the absence of a tractable, parameter-anchored, joint simulation of the training-decay, bounded-control, and adaptive-adversary subsystems with explicit attention to learning-algorithm and feedback-structure robustness. The conservative-lower-bound framing of the adaptive adversary, paired with the dense-sweep functional-form analysis and the cross-algorithm robustness suite, provides a contribution not otherwise present in the prior literature.

Theoretical Framework and Model

Time and State. We model phishing defense as a discrete-time stochastic process. We map one simulation step to one workday of operational exposure dynamics. Under this mapping, a horizon of $T=8000$ steps corresponds to approximately 31 person-years of exposure dynamics, which is a long-run horizon appropriate for studying steady-state residual loss rather than transient post-intervention dynamics. All time-anchored claims in the paper (training half-life, reinforcement interval, decay rate) refer to this workday-per-step mapping. Alternative mappings would require commensurate rescaling of the decay rate and reinforcement interval but would not change the qualitative findings.

Human Subsystem. Each simulation tracks a single representative user with time-varying training state $\tau_t \in [0,1]$. Training decays multiplicatively each step:

$$\tau_{(t+1)} = \lambda \tau_t$$

where $\lambda \in (0,1)$ is the per-step retention rate. Under our calibration, $\lambda=0.98$ corresponds to roughly 50% retention loss within approximately 35 workdays absent reinforcement, in the early-decay range of the Murre and Dros (2015) replication of the Ebbinghaus (1885) forgetting curve. Periodic reinforcement events of magnitude δ raise training:

$$\tau_t \leftarrow \min(1, \tau_t + \delta) \text{ when } t \bmod P = 0$$

where P is the reinforcement interval in steps and δ is the per-event training pulse.

Conditional click probability follows a logistic susceptibility function of message and training state:

$$P(C_t=1) = \sigma(\beta_0 + \beta_q q_t + \beta_p p_t - \beta_\tau \tau_t)$$

where σ is the standard logistic, $q_t \in [0,1]$ is the message quality (a stand-in for grammar, plausibility, lure design, and other content-side properties), $p_t \in [0,1]$ is personalization (a stand-in for target-specific tailoring), and τ_t is the training state at time t . Higher quality and higher personalization raise click probability; higher training reduces it.

Defensive (Technical) Subsystem. Once a user clicks, organizational technical controls (filters, link rewriting, sandboxing, authentication challenges, automated remediation) probabilistically mitigate the attempt. We collapse this layer into a single mitigation probability $g \in [0,1]$: realized loss $L_t = 1$ if a click occurs and the technical control fails (which happens with probability $1-g$), and $L_t = 0$ otherwise. Expected loss conditional on attempt is therefore:

$$E[L_t] = P(C_t=1)(1-g)$$

A control strength of $g=1$ would eliminate loss; $g<1$ leaves a fraction $1-g$ of clicked attempts as realized loss events. We treat $g=0.5$ as a stylized operational midpoint, where modern email filters achieve high detection rates on known and routine phishing patterns, but novel and targeted campaigns evade detection at substantially higher rates (Verizon, 2024). Our sensitivity analysis covers g across $\{0.05, 0.10, \dots, 0.95\}$ plus extremes at 0.01 and 0.99.

Expected per-step loss is strictly positive whenever $g<1$ and $P(C_t)>0$, since $E[L_t] = P(C_t)(1-g)>0$. This is algebraically guaranteed and serves as a sanity check on simulation behavior rather than as a substantive finding. Therefore, the substantive content of the paper is the magnitude and robustness of the steady-state loss under the calibrated parameterization.

Adversarial Subsystem. The attacker selects an action $a_t = (q_t, p_t)$ at each step from a finite action set $A = a_1, \dots, a_K$. Two attacker classes are studied. The static attacker selects uniformly at random from A and does not update. This is not realistic but serves as a non-adaptive baseline against which adaptation is measured. The adaptive attacker is modeled as an epsilon-greedy multi-armed bandit over A . With probability ε , the attacker explores a random action; with probability $1 - \varepsilon$, the attacker exploits the action with the highest current value estimate. Value estimates are updated by sample-mean averaging:

$$Q_{(t+1)}(a) = Q_t(a) + \frac{1}{N_t(a)} (L_t - Q_t(a))$$

where L_t is the realized loss and $N_t(a)$ is the number of times action a has been selected by step t . Reward feedback in the baseline configuration is the realized loss L_t . We also study a click-only-feedback variant in which the attacker observes click occurrence rather than realized (mitigated-or-not) loss. The bandit is a deliberately minimal model of adaptation: it has no contextual reasoning, no model of the defender, no content generation, and no notion of campaign timing or targeting. We adopted this minimal model because our claim is a conservative-lower-bound claim. If persistent residual loss emerges under such weak adaptation, more sophisticated adversaries can only elevate it.

The 3×3 action grid ($q \in \{0.3, 0.6, 0.9\}$, p same) covers the range of plausible message-quality and personalization configurations but does not exhaust them. The β coefficients are calibrated to click rates at the realized states, that is, the action-space centroid ($q = p = 0.6$) and the high corner ($q = p = 0.9$), rather than at the boundary states ($q = 1, p = 1$) that the model never reaches.

Steady-State Residual Loss. We define steady-state residual loss as the average per-step loss over the final quarter of the simulation horizon:

$$L_{ss} = \frac{4}{T} \sum_{t=(3T/4+1)}^T L_t$$

We use this finite-horizon average rather than an analytical equilibrium because the joint dynamics (multiplicative decay plus periodic reinforcement plus stochastic clicks plus bandit updates) do not admit a closed-form fixed point. We use the term “steady-state residual loss” rather than “equilibrium” except when citing economic-equilibrium results (Gordon and Loeb, 2002). The stability of L_{ss} across seeds, parameter sweeps, and learning-algorithm choices substitutes for the formal equilibrium guarantees that a closed-form solution would provide.

Methodology

We organize the paper’s claims as one methodological contribution and one testable hypothesis. We do not advance a hypothesis on diminishing returns to control strength because, absent a cost layer on g , the data show an approximately linear loss versus residual exposure relationship.

Contribution 1 (Methodological: conservative lower bound on adaptive capability). The conservative-lower-bound framing of the adversary subsystem is the paper’s primary substantive contribution. We deliberately model the adversary as a minimally adaptive multi-armed bandit: epsilon-greedy selection over a discrete action space, with reward feedback restricted to realized per-step loss and no contextual reasoning, content generation, or defender modeling. Under this minimal specification, any steady-state loss premium that emerges is bounded below by the model’s logic. More sophisticated adversaries, including contextual bandits, full reinforcement-learning policies, and large-language-model-driven

content generators, can be expected to elevate the floor but not to reduce it. The substantive empirical content is the magnitude and robustness of the floor under the minimally-adaptive specification, examined through cross-algorithm and cross-feedback robustness checks.

Hypothesis H1 (Testable: adaptive premium). Under equivalent defensive parameters, adaptive adversaries (epsilon-greedy bandit) produce strictly higher steady-state loss than static (uniform-random) adversaries. We test this by two-sample t-test comparing baseline-static and baseline-adaptive runs at $g=0.5$, and by reporting the premium across alternative learning algorithms (Thompson, softmax), feedback structures (click-only), and action-space sizes ($K=25$). The robustness of the premium across these variants is the central evidence for Contribution 1.

Experimental Condition. We evaluate four primary experimental conditions and four robustness conditions.

Primary: (1) Baseline static: static (uniform-random) attacker at $g=0.5$, $T=8000$ steps, $N=200$ random seeds. (2) Baseline adaptive: epsilon-greedy bandit attacker ($\epsilon=0.1$) at $g=0.5$, $T=8000$ steps, $N=200$ random seeds. (3) Control sweep static: static attacker over a dense grid $g \in \{0.01, 0.05, 0.10, \dots, 0.95, 0.99\}$ (21 points), $N=50$ seeds per point, $T=8000$ steps. (4) Control sweep adaptive: epsilon-greedy adaptive attacker over the same dense grid, $N=50$ seeds per point.

Robustness: (5) Decay-rate sweep: $\lambda \in \{0.95, 0.98, 0.99\}$ with all other parameters at baseline, $N=100$ seeds. (6) Reinforcement-interval sweep: $P \in \{500, 1000, 2000\}$ with all other parameters at baseline, $N=100$ seeds. (7) Learning-algorithm sweep: Thompson sampling (Beta-Bernoulli posterior); softmax (Boltzmann, temperature $T=0.1$); click-only feedback (epsilon-greedy with reward = click occurrence rather than realized loss); enlarged action space ($K=25, 5 \times 5$ grid). All robustness conditions at $g=0.5$, $T=8000$ steps, $N=100$ seeds.

Outcome Measures and Convergence. The primary outcome is steady-state residual loss L_{ss} as defined above. We report mean, standard deviation, 95% confidence interval, and the seed-level distribution. For control-sweep conditions we additionally fit four functional forms (linear, power-law, log, exponential) by least squares and compare them by Akaike Information Criterion. For the bandit conditions we report the tail-of-run action-selection frequency to identify the most-selected arm.

We assess convergence by inspecting moving-average loss trajectories with window $w=800$ (10% of horizon). Strict per-seed convergence criteria of the type sometimes used in stochastic simulation are not met under this configuration: stochastic oscillation around the steady-state mean is too large for per-seed convergence at standard tolerances. We instead report convergence at the mean-trajectory level. Visual inspection of the moving average across the 200-seed ensemble shows convergence well before step 4,000 in both the static and adaptive conditions, leaving the final $T/4=2,000$ steps as the steady-state estimation window. The stability of L_{ss} across seeds, parameter sweeps, and learning-algorithm choices provides the empirical substitute for per-seed convergence.

Table 1 reports the calibrated parameter values, with a Source column distinguishing parameters anchored to identifiable empirical literature, parameters calibrated to produce literature-consistent realized outcomes at the action-space centroid, and parameters that are stylized choices without a specific empirical anchor. The β coefficients are calibrated to realized states ($q \in 0.3, 0.6, 0.9$, p same), not to boundary states ($q=1, p=1$). Reading the coefficients off the realized centroid produces interpretable and literature-consistent click rates within the 0.10 to 0.40 range.

Table 1. Calibrated Parameter Values, Sources, and Realized-State Implications.

Parameter	Symbol	Value	Source	Implication at action-space centroid
Training decay rate	λ	0.98 (range 0.95-0.99)	Empirical	Approx. 50% retention loss in ~35 workdays, in the early-decay range of Murre and Dros (2015)
Training pulse magnitude	δ	0.3	Calibrated	At centroid $q = p = 0.6$ with $\tau = 0$, induces approximately 17 percentage-point click-rate reduction; consistent with Sommestad and Karlzén (2019) central tendency
Training reinforcement interval	P	1000 (range 500-2000)	Stylized	At one-step-per-workday mapping, sweep covers approximately two to eight years; treated as stylized parameter
Control strength	g	0.5 (sweep 0.01-0.99)	Stylized	Operational midpoint between high-efficacy filtering on known patterns and low-efficacy on novel/targeted campaigns (Verizon, 2024)
Logistic intercept	β_0	-0.5	Stylized	At centroid with $\tau = 0$, baseline click probability ≈ 0.83 ; produces click rates within the 0.10-0.40 range across realized parameter combinations consistent with Sommestad and Karlzén (2019)
Message-quality coefficient	β_q	2.0	Calibrated	Approximately 13 percentage-point lift from low- q (0.3) to high- q (0.9); consistent with spear-phishing-vs-generic effect sizes in Sommestad and Karlzén (2019)
Personalization coefficient	β_p	1.5	Stylized	Lift from low- p (0.3) to high- p (0.9); magnitude larger than small personalization effects in some field meta-analyses, treated as stylized
Training coefficient	β_τ	3.0	Calibrated	At centroid, approximately 9 percentage-point click-rate reduction from $\tau = 0$ to $\tau = 0.5$; consistent with Sommestad and Karlzén (2019)
Attacker exploration	ϵ	0.1	Stylized	Standard epsilon-greedy choice; sensitivity confirmed via Thompson and softmax conditions
Action-space size	K	9 (3×3)	Stylized	Captures realized range of message-quality and personalization; $K = 25$ robustness check confirms qualitative findings

Results

Under the static (uniform-random) attacker at $g=0.5$, the system converges to a steady-state residual loss of $L_{ss}^{\text{static}} = 0.403$ (SD = 0.0097, 95% CI [0.401, 0.404], N = 200 seeds). Under the adaptive epsilon-greedy attacker at the same g , $L_{ss}^{\text{adaptive}} = 0.452$ (SD = 0.0146, 95% CI [0.450, 0.454], N = 200 seeds). The adaptive premium over the static baseline is +12.23%, and the difference is statistically overwhelming (two-sample t-test, $t = 39.7$, $p < 1e-130$). This confirms Hypothesis H1.

Figure 1 shows the moving-average loss trajectories for both conditions on the same axes. We verified that simulated trajectories satisfy the algebraic identity $E[L_t] = P(C_t)(1-g)$ at the $(1-g)=0.5$ ratio throughout the run, confirming correct implementation of the loss-from-click coupling.

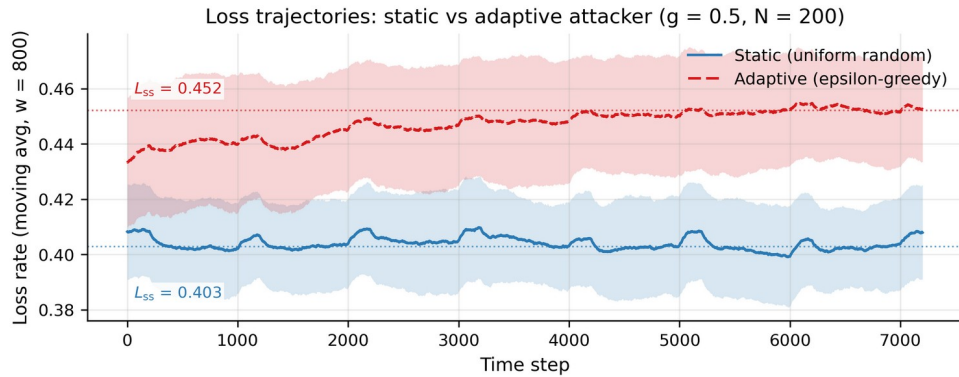


Figure 1. Moving-Average Loss Trajectories Under Static and Adaptive Attackers

Arm Concentration Under Adaptation. The bandit’s tail-of-run action-selection distribution is highly concentrated. The $(q=0.9, p=0.9)$ action captures 56.1% of selections in the last quarter of the simulation, with $(q=0.6, p=0.9)$ capturing an additional 12.8% and $(q=0.9, p=0.6)$ capturing 22.3%. The three actions on the high- q -high- p corner of the grid jointly account for over 91% of selections.

Figure 2 visualizes this distribution as a heatmap. Training decay leaves users persistently susceptible to high-quality, high-personalization messages, and the bandit reliably identifies and exploits this. Under the $K=25$ (5×5 grid) robustness condition, the most-selected arm is again $(q=0.9, p=0.9)$ and captures 42.6% of selections; the dilution relative to $K=9$ reflects only that the grid is finer.

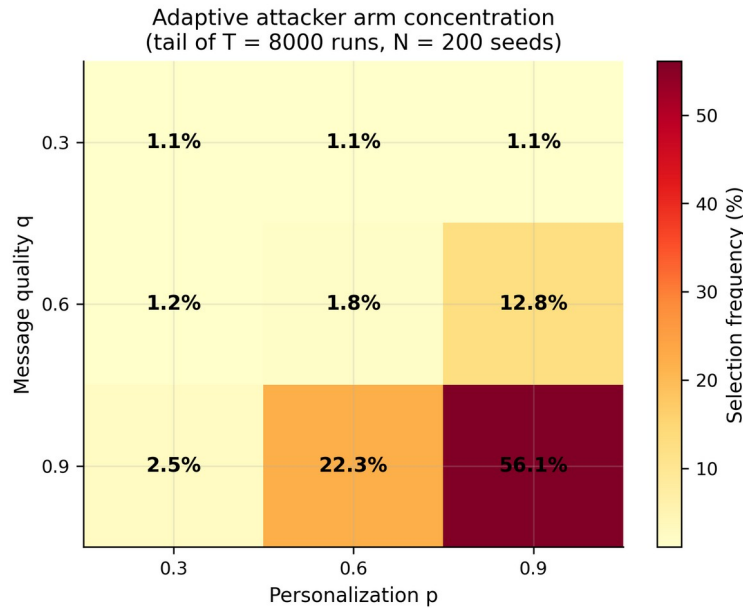


Figure 2. Adaptive Attacker Arm-Selection Frequency in Final

Robustness Across Modeling Choices. The adaptive premium is preserved across learning algorithms and feedback structures. Figure 3 reports the premium across five attacker variants.

Steady-state loss varies in $[0.448, 0.451]$ across $\lambda \in \{0.95, 0.98, 0.99\}$ and in $[0.450, 0.453]$ across $P \in \{500, 1000, 2000\}$. The 95% confidence intervals overlap throughout both sweeps. Under the calibrated click-rate range (0.10 to 0.40), training-cadence variation does not modulate the steady-state floor. This is a result about where defensive leverage lies in the joint system: the floor is set primarily by the bounded-control and adversary-adaptation subsystems; training acts to bound the upper end of click

probability rather than to depress the lower end of the loss floor. Calibrations that deepen the β_τ training coefficient relative to the content coefficients ($\beta_0, \beta_q, \beta_p$) would produce stronger training effects, but the qualitative location of leverage under operational click-rate conditions appears robust.

Across learning algorithms: Thompson sampling produces $L_{ss}=0.453$ (premium +12.58%); softmax produces $L_{ss}=0.423$ (premium + 4.90%); click-only feedback produces $L_{ss}=0.457$ (premium +13.52%); $K = 25$ epsilon-greedy produces $L_{ss}=0.447$ (premium +10.98%). Three of four robustness variants reproduce the +11 to +14 percent premium of the baseline; the softmax variant produces a smaller but still strictly positive premium, interpretable as a property of the exploration-exploitation balance: softmax with temperature $T=0.1$ explores more uniformly than epsilon-greedy with $\epsilon=0.1$, so it concentrates less aggressively on the highest-loss arm and therefore produces less loss elevation.

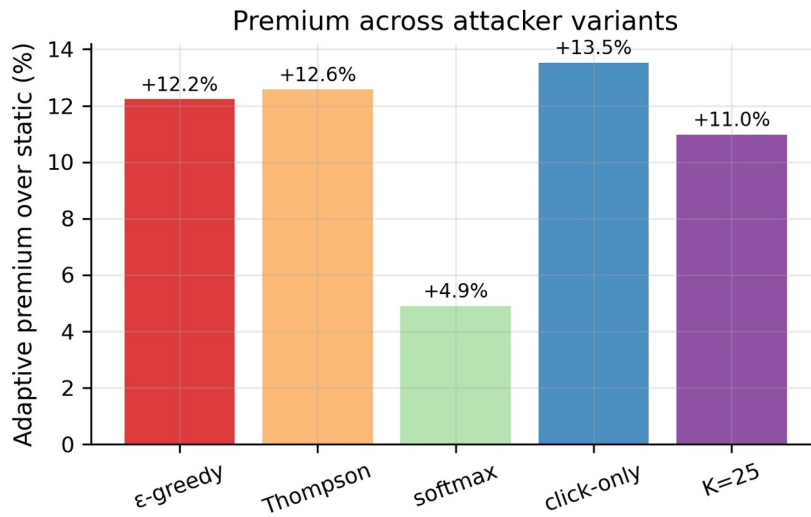


Figure 3. Adaptive Premium Over Uniform-Random Static Baseline Across Five Attacker Variants

Control-Sweep Functional Form. Figure 4 shows the steady-state loss as a function of control strength g across 21 grid points for both static and adaptive attackers. We fit four functional forms by least squares: linear ($L_{ss}=a+bg$), power-law ($L_{ss}=L_0(1-g)^k$), log ($L_{ss}=a+b\log(1-g)$), and exponential ($L_{ss}=a\exp(-bg)$), comparing via the Akaike Information Criterion.

For the static attacker, the power-law fit is preferred over linear by $\Delta AIC=3.3$ ($L_0=0.81, k=1.00$); the k coefficient is essentially one, indicating that the static-attacker loss is statistically indistinguishable from linear in $(1-g)$. For the adaptive attacker, the power-law fit is decisively preferred over linear by $\Delta AIC=31.8$ ($L_0=0.92, k=1.02$). The k coefficient is slightly above one, indicating modestly super-linear curvature; the decisiveness of the AIC preference reflects the dense 21-point sweep and the precision of the per-point estimates ($N=50$ seeds per point).

The cross-attacker exponent comparison ($k_{static} \approx 1.00$ vs $k_{adaptive} \approx 1.02$) is the substantive finding. Adaptive adversaries push the loss-versus-control-strength curve to slightly higher curvature than static adversaries do, because they concentrate selections on the highest-loss arms regardless of g . The absolute magnitude of the curvature is small; the loss reduction per unit of control strength is approximately constant, with a small adaptive-induced curvature that is statistically detectable but not quantitatively dramatic.

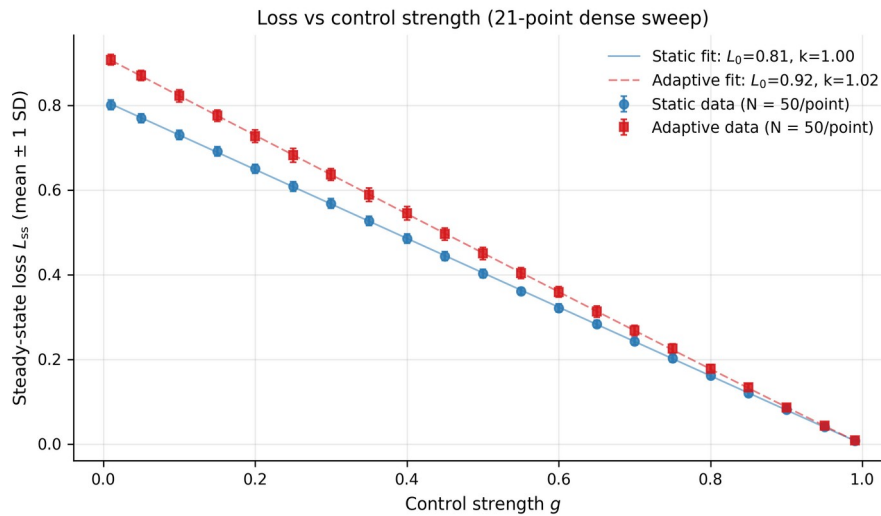


Figure 4. Steady-State Loss Versus Control Strength

Relation to AI-Enabled Threats

The mechanism by which AI-enabled adversaries elevate the steady-state loss floor in this model is visible directly in the $K = 25$ robustness condition. Even when the action space is enlarged from a 3×3 grid to a 5×5 grid, the bandit identifies and concentrates on the high-quality high-personalization corner: the most-selected arm captures 42.6% of selections at $K = 25$, versus 56.1% at $K = 9$, with the dilution explained purely by the finer grid. This concentration pattern is the operational signature of AI-enabled phishing in our framework. Large-language-model-generated content shifts q and p toward their upper values; richer generation capacity expands K ; faster online learning sharpens the bandit's identification of the high-loss corner. Each of these elevates the bandit's mean realized loss, and the cross-algorithm robustness suite shows that this elevation does not depend on the specific learning algorithm.

The model's adversarial subsystem is intentionally minimal: epsilon-greedy bandit selection over a discrete action space, with no content generation, no contextual reasoning, no model of the defender's training cadence, no awareness of post-click outcomes beyond the realized-loss reward. This minimality is deliberate. It establishes a conservative lower bound: if epsilon-greedy adaptation already produces a +12% premium over uniform-random selection, more capable adversaries, including AI-enabled ones, can only raise the floor.

Recent work demonstrates that large language models qualitatively change the adversarial-content landscape. Heiding et al. (2024) report a controlled experiment in which large-language-model-generated phishing emails achieved click rates statistically comparable to expert-crafted emails, and human recipients were unable to reliably distinguish AI-generated from legitimate communications. The marginal cost per high-quality phishing email approaches zero under generative tooling, so the practical constraint that historically limited spear-phishing to high-value targets is loosened.

In our model, the AI-enabled-threats scenario maps to four parameter movements: higher message quality q (large language models produce grammatically perfect, contextually plausible copy); higher personalization p (generative models combined with public-data scraping can tailor each message to the recipient's role, recent activity, and social context); larger effective action space K (AI-enabled adversaries

can generate orders-of-magnitude more strategy variations than human attackers, and the $K = 25$ robustness condition demonstrates that the bandit identifies and concentrates on the high- q high- p corner even with richer action spaces); and faster effective adaptation (richer policies such as Thompson sampling, contextual bandits, or full reinforcement learning would tighten the rate at which the adversary discovers high-loss strategies, and the Thompson-sampling robustness check confirms that the floor is robust to algorithmic strengthening).

Taken together, these mappings imply that AI-enabled adversaries are not a categorical break with our model; they are an extreme corner of the parameter space the model already covers. Under the model, the predicted effect of an AI-assisted adversary is an upward shift in the q and p marginals of the bandit's selection distribution and, through that shift, an elevation in the steady-state loss floor. We are cautious about quantitative magnitude: the β coefficients are calibrated to realized states observed in the simulation, not to boundary states ($q=1, p=1$) that would be more representative of AI-enabled content, and extrapolating to boundary states produces click probabilities that exceed the literature-anchored ranges. The directional claim, that AI-enabled adversaries elevate the floor, is licensed by the model's structure; precise magnitude estimates are not.

The takeaway for practitioners is that the conservative-lower-bound floor we report is a baseline against which AI-enabled adversaries should be evaluated, not a ceiling. Organizations planning for an AI-augmented adversarial landscape should expect the floor to rise, and the substantive question becomes how much, not whether.

Discussion

Three substantive findings emerge from the simulation. The first is that residual loss is structurally positive across modeling choices. The algebraic component is trivial (it follows from $E[L_t] = P(C_t)(1-g) > 0$ whenever $P(C_t) > 0$ and $g < 1$); the substantive component is that the magnitude L_{ss} in the 0.40 range under the baseline calibration is sufficient to drive non-trivial expected per-step incident rates over the simulation horizon. The contribution lies in the magnitude and the robustness of that magnitude, not in the existence of positive loss. The second is that the adaptive premium is robust at approximately 11 to 14%. Across epsilon-greedy, Thompson sampling, click-only feedback, and $K=25$ conditions, the premium falls in a narrow [+10.98%, +13.52%] band. Softmax produces a smaller but still positive premium. The premium is not an artifact of any one learning algorithm or feedback structure; it is a property of the joint defender-attacker dynamics under any adversary that concentrates selections on high-loss arms. The third is that concentration on high-quality high-personalization arms is reliable. Under $K=9$, 56.1% of selections in the steady-state window land on ($q=0.9, p=0.9$); under $K=25$, the corresponding figure is 42.6%. The concentration is what makes the adaptive premium possible: training decay leaves the user susceptible primarily to high- q -high- p messages, and the bandit identifies this. We describe this as a decay-driven concentration chain in the cross-subsystem-interaction sense.

These three findings taken together illustrate cross-subsystem interactions that any single-subsystem analysis would miss. The training-decay literature alone cannot predict the concentration pattern; it would observe that susceptibility decays but would not engage with how attackers respond. The bounded-controls literature alone cannot predict the adaptive premium; it would observe that some fraction of clicked attempts evade mitigation but would not engage with the attacker's strategy selection. The adversarial-adaptation literature alone cannot predict the magnitude of the loss floor; it would observe that adaptation produces strategy concentration but would not engage with the time-varying user state that makes some strategies

more profitable than others. The conservative-lower-bound framing connects these three threads: under the weakest plausible adversary model, the joint dynamics already produce a measurable, robust loss premium. Three practitioner implications follow. First, the steady-state loss floor is the correct planning baseline for any organization evaluating exposure to adaptive phishing campaigns. Under our calibration the floor sits in the 0.40 to 0.45 range at moderate control strength; the cross-algorithm robustness suite means an organization should expect a floor of comparable magnitude regardless of which specific learning algorithm an adversary deploys. The floor, not the static-attacker baseline, is the operational planning anchor. Second, the leverage analysis implies that training-cadence investment alone is bounded in its capacity to depress the floor under operational click-rate conditions; resource allocation that treats training cadence as the primary lever will see returns bounded above by the structural floor set by adversarial concentration and control strength. The dominant leverage in the model resides in the control-strength and adversary-mitigation subsystems. Third, the conservative-lower-bound framing converts the practitioner question “how much residual loss should I plan for under AI-enabled phishing” from speculation to a structured analysis: the floor we report is the baseline; AI-enabled adversaries elevate but do not invalidate it.

We are deliberately cautious about three classes of claim sometimes advanced in the security-economics literature that are not honestly defensible from this simulation as currently specified. First, we resist the framing of phishing defense as an optimization problem rather than a prevention problem. The data establish a steady-state loss floor; they do not establish what optimal investment looks like, because that requires a cost layer we do not include. Second, we resist strong “diminishing returns to control strength” claims, because under the calibrated parameters the data show near-linear loss reduction in $(1 - g)$. Third, we resist managerial claims that conflate model implications with empirical recommendations: our coefficients are stylized and calibrated to literature ranges, not fitted to organizational data, so the magnitudes we report are simulation outputs rather than operational predictions.

Limitations and Future Research

Model and calibration. Under the calibrated parameters, training-cadence variation produces small effects on steady-state loss because residual click probability remains well above zero across the cadence range; a different calibration (lower β_0 , equivalent recalibration of β_q and β_p) would produce stronger training-cadence effects. Per-seed convergence at strict tolerances is not achieved because stochastic oscillation around the mean is too large; we report mean-trajectory convergence across the ensemble instead. The β coefficients and the λ, P, δ values are anchored to literature ranges (Sommestad and Karlzén, 2019; Murre and Dros, 2015) but not fitted to organizational phishing-simulation logs; empirical calibration would be required to convert the simulation from explanatory to predictive.

Adversary model. The action space is small relative to real phishing campaigns. The $K = 9$ and $K = 25$ grids are stylized discretizations of the continuous (q, p) space, and a richer action space with time-varying message features or seasonal targeting is not represented. The static (uniform-random) attacker is an artificial baseline; real adversaries do not select strategies uniformly at random, so the +12% adaptive premium is a lower bound on the premium of a sophisticated adversary over a realistic-static adversary.

Economic and structural extensions. The model tracks a single representative user, which isolates the joint subsystem dynamics but does not capture organization-level heterogeneity. Extending the single user to a heterogeneous population, with role-, tenure-, and department-specific susceptibility and training response, would let the model speak to the distribution of loss across an organization rather than the expected loss of one user. Calibrating that population to real organizational data, as described above, is a priority for future work, alongside multi-step-campaign and stateful-adversary extensions.

Conclusion

We develop a discrete-time simulation that couples training-decay, bounded-control, and adaptive-adversary subsystems within a unified socio-technical framing. The paper's primary contribution is methodological: by modeling the adversary as a minimally adaptive multi-armed bandit, we establish a conservative-lower-bound framing in which residual phishing loss is a structural property of the joint defender-attacker dynamics rather than an artifact of insufficiently aggressive component-level investment. Under that framing, our empirical results, robust across alternative learning algorithms, feedback structures, and action-space sizes, establish a floor against which more sophisticated adversaries can be evaluated. What the paper does not establish is also worth naming. It does not establish optimal defensive investment, because the cost layer required for that claim is reserved for a journal-length extension. It does not establish empirical predictiveness, because the parameterization is calibrated to literature ranges rather than fitted to organizational data. It does not, on its own, license claims about the magnitudes of specific AI-enabled-elevation predictions; the calibration restricts those to qualitative directional claims. What it does establish is that, under the joint dynamics of three independently-studied subsystems, the steady-state residual loss floor is robust and measurable, that minimal adversarial adaptation elevates it in a manner that does not depend on the specific learning algorithm, and that this floor is the appropriate baseline against which AI-enabled scenarios should be evaluated.

Acknowledgements: The author used generative artificial-intelligence tools during the preparation of this manuscript. Specifically, such tools assisted with some drafting and editing, with review of the simulation code and analysis, and with literature summarization. All AI-assisted material was reviewed, verified, and edited by the author, who takes full responsibility for the content of the paper. This disclosure is provided in accordance with the IIS AI Use Policy (https://iacis.org/iis/AI_Policy.php).

References

- Abbasi, A., Dobolyi, D., Vance, A., and Zahedi, F. M. (2021). The phishing funnel model: A design artifact to predict user susceptibility to phishing websites. *Information Systems Research*, 32(2), 410-436.
- Beautement, A., Sasse, M. A., and Wonham, M. (2008). The compliance budget: Managing security behaviour in organisations. *Proceedings of the 2008 New Security Paradigms Workshop*, 47-58.
- Biggio, B., and Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84, 317-331.
- Bostrom, R. P., and Heinen, J. S. (1977). MIS problems and failures: A socio-technical perspective. *MIS Quarterly*, 1(3), 17-32.
- Burns, A. J., Posey, C., Courtney, J. F., Roberts, T. L., and Nanayakkara, P. (2017). Organizational information security as a complex adaptive system: Insights from three agent-based models. *Information Systems Frontiers*, 19(3), 509-524.
- Canfield, C. I., Fischhoff, B., and Davis, A. (2016). Quantifying phishing susceptibility for detection and behavior decisions. *Human Factors*, 58(8), 1158-1172.

- Caputo, D. D., Pfleeger, S. L., Freeman, J. D., and Johnson, M. E. (2014). Going spear phishing: Exploring embedded training and awareness. *IEEE Security and Privacy*, 12(1), 28-38.
- Ebbinghaus, H. (1885). *Memory: A contribution to experimental psychology* (H. A. Ruger and C. E. Bussenius, Trans.). Teachers College, Columbia University.
- Gordon, L. A., and Loeb, M. P. (2002). The economics of information security investment. *ACM Transactions on Information and System Security*, 5(4), 438-457.
- Grossklags, J., and Johnson, B. (2009). Uncertainty in the weakest-link security game. *Proceedings of the 2009 International Conference on Game Theory for Networks*, 673-682.
- Heiding, F., Schneier, B., Vishwanath, A., Bernstein, J., and Park, P. S. (2024). Devising and detecting phishing emails using large language models. *IEEE Access*, 12, 42131-42146.
- IBM Security. (2024). Cost of a data breach report 2024. IBM Corporation.
<https://www.ibm.com/reports/data-breach>
- Jensen, M. L., Dinger, M., Wright, R. T., and Thatcher, J. B. (2017). Training to mitigate phishing attacks using mindfulness techniques. *Journal of Management Information Systems*, 34(2), 597-626.
- Laszka, A., Felegyhazi, M., and Buttyan, L. (2014). A survey of interdependent information security games. *ACM Computing Surveys*, 47(2), 1-38.
- Murre, J. M. J., and Dros, J. (2015). Replication and analysis of Ebbinghaus' forgetting curve. *PLOS ONE*, 10(7), e0120644.
- National Institute of Standards and Technology. (2024). Building a cybersecurity and privacy learning program: NIST Special Publication 800-50 revision 1 (SP 800-50r1).
<https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-50r1.pdf>
- Nguyen, T. T., and Reddi, V. J. (2023). Deep reinforcement learning for cyber security. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8), 3779-3795.
- Reason, J. (2000). Human error: Models and management. *BMJ*, 320(7237), 768-770.
- Sasse, M. A., and Flechais, I. (2005). Usable security: Why do we need it? How do we get it? In L. F. Cranor and S. Garfinkel (Eds.), *Security and usability* (pp. 13-30). O'Reilly Media.
- Sommestad, T., and Karlzén, H. (2019). A meta-analysis of field experiments on phishing susceptibility. *Proceedings of the 2019 APWG Symposium on Electronic Crime Research (eCrime)*, 1-14.
- Steves, M., Greene, K., and Theofanos, M. (2020). Categorizing human phishing difficulty: A Phish Scale. *Journal of Cybersecurity*, 6(1), tyaa009.
- Tambe, M. (2011). *Security and game theory: Algorithms, deployed systems, lessons learned* (1st ed.). Cambridge University Press.

Verizon. (2024). 2024 data breach investigations report. Verizon Enterprise Solutions.
<https://www.verizon.com/business/resources/reports/dbir/>