



Research Article

ClickGuard: An Experimental Field Study of Phishing Simulation Training Effectiveness in an Organizational Setting

Aminu Muhammad Auwal^{1*}  and Sulaiman Shehu² 

¹Faculty of Natural Sciences, University of Jos, Plateau State, Nigeria

²Department of Software Engineering, Bayero University, Kano State, Nigeria

Article Information

Abstract

Article History

Received: 25 June 2026

Revised: 26 July 2026

Accepted: 20 August 2026

Published online: 30 September 2026

Keywords

Phishing Simulation
Security Awareness Training
Field Experiment
Click-Through Rate
Reporting Rate
Social Engineering
Authority Cues
Organizational Security
Cybercrime
Embedded Training

Correspondence*

i.elameenu@gmail.com

ORCID

Aminu Muhammad Auwal
<https://orcid.org/0009-0005-1799-7876>
Sulaiman Shehu
<https://orcid.org/0009-0002-0056-604X>

Simulated phishing campaigns with embedded training are widely used in organizational cybersecurity programs, yet evidence regarding their long-term effectiveness remains mixed. Questions also persist regarding their ability to address evolving phishing techniques and differing patterns of workforce susceptibility. This study evaluates the effectiveness of embedded micro-training and video-based interventions in reducing phishing susceptibility and improving reporting behavior. This study evaluates whether a combined embedded micro-training and video module intervention produces significantly greater reductions in phishing click-through rates (CTR) and improvements in phishing reporting rates compared with standard security policy reminders delivered to a control group. A secondary objective is to determine whether these effects differ across three message themes exploiting authority, scarcity, and generic delivery cues, enabling a theme-specific assessment of training effectiveness that aggregate CTR figures cannot provide. An 8-week randomized field experiment was conducted within a mid-sized organization of 612 employees. Employees were stratified by department and randomly assigned to a training group ($n = 306$) or control group ($n = 306$). Three simulated phishing campaigns were delivered at Weeks 1, 4, and 8, each using a distinct social engineering theme. Training group members who clicked a simulated phishing link received immediate embedded micro-training plus a 10-minute structured video module. Control group members received only the standard annual security policy reminder. The training group's CTR fell from 28.3 percent at baseline to 14.7 percent post-intervention, compared with a decline from 27.1 percent to 22.4 percent in the control group (group-by-campaign interaction $p < 0.001$). Reporting rate increased from 9.8 percent to 31.2 percent in the training group versus 10.5 percent to 14.1 percent in the control group. Median time to report fell from 142 minutes to 47 minutes in the training group. Authority-themed messages showed substantially smaller training-related improvements than generic or scarcity-themed messages, with the training group's authority-message CTR remaining 22.6 percent at Week 4 versus 25.8 percent in the control group. Combined embedded micro-training and structured video delivery produce statistically and operationally significant reductions in phishing CTR and improvements in reporting behavior. These effects are theme-dependent: authority-based social engineering is more resistant to brief training interventions and requires supplementary role-specific and technical defenses, including DNS and IP filtering. Sustained organizational investment in cybersecurity education and attention to emerging AI-generated phishing threats are prerequisites for maintaining training effectiveness over time.

© 2026 Centre for Research and Innovation (CRI). This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

I. INTRODUCTION

Phishing remains the dominant initial access technique exploited in organizational data breaches, and its continued effectiveness in the face of substantial technical countermeasures reflects a fundamental asymmetry: attackers need only identify one susceptible user among thousands of email recipients, while defenders must protect

every recipient simultaneously. Simulated phishing campaigns have been adopted by organizations worldwide as a mechanism for measuring this susceptibility and for delivering just-in-time training to those who click simulated malicious links. The logic of simulation-based training is theoretically sound: embedding a training intervention at the precise moment of susceptibility demonstrated behavior should produce stronger learning than generic awareness

content delivered months before or after an individual's actual vulnerability is expressed. However, the empirical evidence on whether this logic translates into durable, operationally significant behavior change is surprisingly ambiguous.

Controlled field experiments have produced divergent findings. A controlled experiment conducted at an Israeli financial institution with approximately 5,000 employees demonstrated that the framing and personalization of simulated phishing emails, rather than the timing of training relative to the campaign, were the dominant determinants of CTR, leading to recommendations for data-driven, tailored simulation design rather than reliance on a single generic template [1]. In sharp contrast, a large-scale randomized controlled experiment at a major healthcare organization involving over 19,500 employees across ten simulated campaigns over eight months found no statistically significant relationship between recent completion of embedded or annual training and the probability of failing a subsequent phishing simulation, with most employees spending minimal time engaging with embedded training material after clicking [2]. These divergent findings are difficult to reconcile because the studies differed on multiple dimensions simultaneously, including organizational size, sector, training content, campaign frequency, and outcome metrics, making it difficult to identify which features of training program design are responsible for effectiveness differences.

The present study addresses this interpretive gap by conducting a focused 8-week randomized field experiment within a single organization, deliberately holding constant the features that varied across prior studies, including training content, campaign frequency, and message themes, while randomizing exposure to the training intervention. This design enables clean causal identification of the training effect and supports theme-level analysis of whether training effectiveness varies across different social engineering cue types. The study is further motivated by the emerging challenge of adversarially generated phishing content: research on adversarial threat modeling in generative AI has identified spear-phishing message generation as a high-priority attack vector in which large language models can produce highly convincing, contextually tailored phishing messages at scale, potentially rendering training programs calibrated against conventional phishing templates less effective against the next generation of attacks [3]. At the same time, technical controls such as DNS and IP filtering provide a complementary upstream defense layer that interacts with training effectiveness in ways that are rarely modeled explicitly in experimental studies [4]. Closely related field evidence from three organizations indicates that the effectiveness of embedded phishing training varies by lure type and employee engagement [21], and survey evidence on ransomware social-engineering lures identifies workplace time pressure and personality traits as correlates of susceptibility [22].

The sociological and criminological dimensions of organizational phishing susceptibility are also relevant context. Auwal and Lazarus [5] documented through ethnographic methods how socioeconomic factors and lifestyle patterns shape victimization risk among targets of Nigerian Yahoo Boys, and their broader criminological categorization framework [6] provides a foundation for situating organizational phishing risk within a wider victimization typology that goes beyond individual psychological traits. Organizational security culture, the shared norms and values that govern how employees approach security decisions, represents a collective analog of individual security knowledge, and organizations with underdeveloped security cultures may show systematically higher baseline susceptibility and weaker training responses than those with mature security governance frameworks. The ethical dimensions of deploying simulated attacks on employees, including questions of proportionality, consent, and data use, require explicit governance frameworks that are not always in place, with consequences for both program legitimacy and employee trust [7]. Analysis of Nigerian EFCC conviction records further shows that cybercrime enforcement is concentrated around impersonation-based and cyber-affiliated offences [23], which is relevant context for organizations facing impersonation-style phishing lures.

This paper makes four principal contributions: (i) a pre-registered, stratified randomized field design that cleanly isolates the combined effect of embedded micro-training and structured video instruction from simulation exposure alone; (ii) simultaneous reporting of CTR and reporting rate as complementary outcome measures that capture different dimensions of security behavior improvement; (iii) theme-level analysis across authority, scarcity, and generic delivery cues that reveals heterogeneity in training effectiveness that aggregate CTR figures conceal; and (iv) a contextually grounded discussion situating these findings within the broader landscape of AI-generated phishing threats, technical countermeasures, and ethical governance requirements.

II. RELATED WORK

A. The Organizational Phishing Simulation Literature

Field experiments embedded within real organizations provide the most ecologically valid evidence on phishing training effectiveness because they capture behavior in employees' normal working environments, under realistic time pressures and alongside the distractions of actual job responsibilities. The Israeli financial institution study [1] is notable for its large scale and for the sophistication of its experimental design, which varied both message personalization and training delivery timing within a factorial structure. The finding that personalization, specifically the use of the recipient's name and role-relevant content in the phishing message, substantially elevated CTR is theoretically important because it implies that generic training programs that prepare employees to recognize generic-template phishing messages may not generalize to personalized or

targeted spear-phishing attacks. The study's recommendation for a data-driven approach to simulation campaign design, in which historical click data are used to select and refine message templates, anticipates the adversarial AI challenge by pointing toward continuous adaptation of training stimuli.

Williams, Hinds, and Joinson [8] conducted one of the largest published field studies of workplace phishing susceptibility, exposing over 60,000 employees across multiple organizations to nine spear-phishing simulation campaigns over six weeks. Their principal finding, that authority-framed messages (those appearing to come from a senior organizational figure) produced substantially higher CTRs than urgency- or curiosity-framed messages, and that susceptibility varied significantly across organizational roles, has important implications for theme-specific training design. Client-facing and administrative roles were most susceptible to authority-framed messages, likely because these roles involve high volumes of legitimate requests from senior personnel, creating a context in which the authority compliance heuristic is regularly reinforced rather than critically examined. The present study's observation of a similar authority-theme effect at the campaign level provides field replication of this pattern in an independent organizational sample.

The large-scale healthcare organization experiment [2] remains the most rigorous published test of embedded training effectiveness and the most challenging finding for proponents of simulation-based training programs. The study's null finding, that embedded training and annual awareness training alike did not significantly reduce the probability of clicking a subsequent simulated phishing message, is particularly striking because it was obtained over an 8-month horizon with ten campaigns, a design with substantially more statistical power than most prior studies. The authors suggest several explanations: that employees spend too little time engaging with embedded training content to encode meaningful behavior change; that the training content was not sufficiently tailored to the specific message types employees encountered; and that simulation frequency creates habituation effects in which employees who have been exposed to multiple campaigns begin to treat simulation emails as a known category rather than practicing genuine threat evaluation. The present study addresses the first two of these proposed mechanisms by combining triggered embedded micro-training with a structured video module, and by using message themes drawn from documented organizational threat intelligence.

B. Cybercrime Victimization in Organizational Contexts

The organizational phishing training literature has developed largely in isolation from the broader cybercrime victimization literature, which offers theoretical resources and empirical findings that are directly relevant to understanding why training effects vary across individuals, roles, and organizations. Routine activity theory [9] posits that victimization risk reflects the intersection of motivated

offenders, suitable targets, and absent guardianship; in the organizational context, capable guardianship is jointly constituted by technical defenses, security policy enforcement, and the behavioral vigilance of individual employees. Training programs can be understood as interventions aimed at strengthening the behavioral component of guardianship, and their effectiveness is contingent on the baseline level of motivation and capability that employees bring to the training task.

The sociological and criminological framing developed by Auwal and Lazarus [5], [6] adds nuance by situating phishing within the spectrum of social engineering techniques deployed by organized cybercriminal networks that deliberately calibrate message framing to target identified vulnerabilities in specific populations. In organizational settings, the most relevant population vulnerabilities include role-based exposure to specific message types, security culture strength, and the socioeconomic pressures that may lead employees to act on urgent or reward-promising messages without full deliberate evaluation. The finding in the present study that authority-themed messages resisted training more effectively than other themes can be partly understood through this lens: authority compliance is a deeply socialized norm in hierarchical organizational cultures, and brief training interventions that present factual information about authority-based phishing may not be sufficient to override ingrained deference to apparent senior-level requests.

The ideal victim concept [10] has also been applied to organizational phishing contexts: employees who click simulated phishing links are sometimes perceived as having failed a test rather than having been deceived, which can generate stigma effects that reduce willingness to report real phishing encounters. The significant improvement in reporting rate observed in the training group in the present study, from 9.8 percent to 31.2 percent, is therefore not merely a secondary outcome but a primary indicator of whether training has shifted employees' understanding of phishing from a personal failure to a security incident requiring collective organizational response.

C. Emerging Threats: Adversarial AI and Technical Control Interactions

The threat landscape that organizational training programs must prepare employees to navigate is evolving rapidly, and in ways that may systematically reduce the generalizability of training effects observed against conventional phishing templates. Auwal [3] conducted a systematic analysis of adversarial threat vectors in generative AI systems, identifying spear-phishing content generation, voice synthesis for vishing attacks, and contextually tailored business email compromise as high-priority attack surfaces in which large language model capabilities dramatically lower the expertise barrier for producing convincing deceptive content. Training programs that use static message templates drawn from past campaigns may be less effective

against AI-generated phishing that is dynamically tailored to each recipient's linguistic style, organizational relationships, and recent online activity. Technical controls provide a complementary layer of defense that interacts with training effectiveness in important ways. DNS and IP filtering systems [4] can block access to known phishing infrastructure at the network level, reducing the volume of successful phishing attempts that bypass user behavior. This means that the behavioral improvement required from training programs is lower in organizations with strong technical controls, potentially making marginal training effects statistically detectable in well-defended organizations that would be undetectable in poorly defended ones. Web application vulnerabilities, including directory brute-forcing and artifact exposure documented by Auwal [11], create lateral escalation pathways that transform a single phishing compromise into a broader organizational breach, making the prevention of initial phishing clicks an increasingly high-stakes behavior in organizations with vulnerable web infrastructure. SMEs with limited technical resources face compounding risks from both dimensions [12], reinforcing that training program design cannot be divorced from the technical security context in which it operates. Internal threat pathways and ethical security lapses in Nigerian IoT deployments [24], together with the technical limitations of IoT-based telemedicine monitoring and data transmission [25], show that these exposures extend to connected systems beyond email.

The ethical dimensions of deploying simulated phishing campaigns add further complexity. Blaise *et al.* [7] documented significant inconsistencies in the ethical frameworks that information security professionals apply when designing and overseeing monitoring programs, and these inconsistencies extend to the conduct of phishing simulation campaigns. Organizations that deploy simulations without transparent governance policies, without ethical review of training content, and without clear policies on how click data are used may undermine the organizational trust that training programs depend on for their effectiveness. The student and employee perspective on AI-driven learning experiences, examined by Auwal [13], suggests that individuals are sensitive to the perceived fairness and transparency of automated or algorithmically managed training interactions, a sensitivity that is likely to extend to the experience of being tested through simulated attacks. Evidence on privacy fatigue and ChatGPT adoption among Nigerian higher education users [26] similarly suggests that attitudes toward data protection shape how individuals engage with technology-mediated safeguards.

D. Psychological Predictors of Training Response

Individual-level psychological research has identified a set of traits and cognitive characteristics that predict not only baseline phishing susceptibility but also the magnitude of training-induced behavior change. Research on information processing models of phishing vulnerability [14] suggests that training is most effective when it interrupts the habitual,

low-elaboration processing pathway that is the proximate cause of susceptibility for most individuals: factual knowledge about phishing indicators produces limited behavior change if users do not shift from peripheral to central processing of incoming email during real threat exposure. This theoretical prediction implies that training programs should include components that practice the application of knowledge under realistic time and cognitive load conditions, rather than simply delivering declarative content. Research on recent phishing encounter effects [15] demonstrates that prior simulation exposure shapes subsequent susceptibility in complex ways. Near-miss experiences, in which a user clicks a simulated link but is then shown an embedded training page before entering credentials, tend to produce short-term improvements in detection performance but may also generate overconfidence in users who believe they have now learned to identify phishing, reducing the vigilance applied to subsequent messages. Detecting the evolving characteristics of phishing attacks, including AI-generated content that closely mimics legitimate communications, requires ongoing and adaptive vigilance rather than a fixed pattern-matching approach, a distinction that most current training programs do not adequately emphasize [16].

III. EXPERIMENTAL DESIGN

A. Study Context and Ethical Governance

The study was conducted within a mid-sized organization employing 612 staff across administrative, technical, client-facing, and management functions. Organizational participation was approved by the organization's information security governance committee and the university research ethics committee. All employees were informed, through the organization's existing security policy communications, that periodic phishing simulations form part of the organization's security awareness program. No disclosure was made of the specific study timeline, campaign design, or group allocation, consistent with standard practice in simulation-based phishing research and in accordance with the ethical governance principles discussed in Blaise *et al.* [7]. Anonymized click and reporting logs were provided to the research team by the organization's security operations team and were handled under a formal data sharing agreement. Individual employee click data were not shared with the organization's management in identifiable form, a governance decision designed to mitigate the stigma effects associated with simulation failures and to encourage honest reporting behavior.

B. Randomization and Balance

Employees were stratified by department and then randomly allocated within each stratum to the training group ($n = 306$) or the control group ($n = 306$) using a computer-generated random allocation sequence. Stratification ensured that departments with distinct role compositions, and therefore distinct baseline susceptibility profiles by message theme,

were proportionally represented in each group. Balance checks at baseline confirmed that the two groups were similar in gender distribution (45.3 percent male in training vs. 46.4 percent in control), age distribution (mean age 34.2 in training vs. 34.7 in control), and departmental composition (no department differed by more than 2 percentage points between groups). Baseline CTRs confirmed group comparability prior to intervention (see Results).

C. Campaign Design and Message Themes

Three simulated phishing campaigns were delivered over the 8-week study period. The Week 1 baseline campaign used a generic delivery notification theme: a message appearing to come from a courier company notifying the recipient that a parcel required address confirmation via an embedded link. The Week 4 interim campaign used an authority theme: a message appearing to originate from a senior executive requesting urgent review of a confidential document accessible via an embedded link. The Week 8 post-intervention campaign used a scarcity theme: a notice informing the recipient that they had been selected to receive an organizational benefit that must be claimed within 24 hours via an embedded link. Message texts were developed by the research team based on real organizational phishing reports and were reviewed for realism by two members of the organization's information security team who were blinded to the experimental design. All three messages were sent from plausible but non-organizational domains using display names consistent with the ostensible sender. Embedded links directed clickers to a landing page hosted by the research team that displayed either the embedded training content (for training group members) or a neutral notification that the message was part of a security awareness exercise (for control group members). Link clicks, report-phishing button presses, and time-to-report values were recorded automatically by the research team's campaign infrastructure.

D. Training Intervention

Training group employees who clicked a simulated phishing link in any campaign were immediately redirected to an embedded micro-training page tailored to the specific social engineering cues present in that campaign's message. The micro-training page presented the original phishing message with annotated highlights identifying the suspicious elements, followed by a brief quiz (three items) assessing recognition of those elements and a set of decision rules for evaluating future messages of similar type. Employees were asked to complete the 10-minute structured video module within 5 business days of clicking; completion was verified through the learning management system. Training group

employees who did not click any link in the baseline or interim campaigns received the video module as a proactive refresher delivered between the baseline and interim campaigns, ensuring that all training group members received the intervention regardless of whether they clicked. The video module covered three content areas: (i) an overview of the four primary social engineering cue categories (authority, urgency, scarcity, curiosity) with real-world examples of each; (ii) a behavioral decision framework for evaluating incoming email, including a step-by-step guide to checking sender authenticity, link destinations, and message content for internal consistency; and (iii) organizational reporting procedures, emphasizing that reporting suspected phishing is a valued security behavior rather than an admission of error, and that the organization's security team relies on employee reports to detect real phishing campaigns. This framing was designed to address the stigma effects associated with simulation failures and to increase reporting rate as a secondary outcome.

E. Outcome Measures and Analysis

The primary outcome measure was CTR: the proportion of campaign recipients in each group who clicked the embedded link in each simulated phishing email. The secondary outcome measures were reporting rate, defined as the proportion of recipients who pressed the report-phishing button before or instead of clicking, and median time to report, defined as the elapsed time in minutes from email delivery to report button press for employees who reported. Differences in CTR and reporting rate between groups and across campaigns were analyzed using chi-square tests for each campaign comparison and logistic regression for the full 3-campaign dataset, with group, campaign, and message theme entered as fixed effects and their two-way interactions included to test for theme-specific and temporal heterogeneity. Median time to report was compared using Wilcoxon rank-sum tests due to the right-skewed distribution of response times.

IV. RESULTS

A. Baseline Comparability and Campaign-Level CTR

Baseline CTRs were 28.3 percent (training group: 87 of 306 employees clicked) and 27.1 percent (control group: 83 of 306 employees clicked), a difference of 1.2 percentage points that was not statistically significant (chi-square = 0.14, df = 1, $p = 0.71$), confirming adequate randomization. Table 1 presents CTR for all three campaigns separately for each group.

TABLE 1 CLICK-THROUGH RATE (PERCENTAGE OF GROUP) BY CAMPAIGN AND GROUP

Campaign	Theme	Training CTR (%)	Control CTR (%)	Difference (pp)	Chi-sq	p
Week 1	Generic delivery	28.3	27.1	1.2	0.14	0.710
Week 4	Authority	22.6	25.8	3.2	1.87	0.171
Week 8	Scarcity	14.7	22.4	7.7	9.63	0.002

The CTR gap between the training and control groups widened monotonically across the three campaigns, from 1.2 percentage points at baseline to 7.7 percentage points at the post-intervention assessment, a significant group-by-campaign interaction in logistic regression (odds ratio for group-by-campaign interaction = 0.72 per campaign, 95 percent CI 0.61 to 0.85, $p < 0.001$). The authority-themed campaign at Week 4 showed a smaller differential (3.2 pp) than the scarcity-themed campaign at Week 8 (7.7 pp), and

the Week 4 difference was not statistically significant ($p = 0.171$), suggesting that the training intervention had not yet produced a meaningful reduction in authority-cue susceptibility by the midpoint of the study. The Week 8 difference was statistically significant ($p = 0.002$), with the training group's CTR for the scarcity-themed message having declined by 13.6 percentage points from baseline while the control group's CTR declined by only 4.7 percentage points.

B. Reporting Rate and Time to Report

TABLE II REPORTING RATE (%) AND MEDIAN TIME TO REPORT (MINUTES)

Campaign	Training Reporting (%)	Control Reporting (%)	Difference (pp)	Training Median (min)	Control Median (min)
Week 1 Baseline	9.8	10.5	0.7	142	148
Week 4 Interim	21.4	12.7	8.7	88	139
Week 8 Post-Intervention	31.2	14.1	17.1	47	136

Reporting rate in the training group increased by 21.4 percentage points from baseline to the post-intervention campaign (9.8 percent to 31.2 percent), while the control group's reporting rate increased by only 3.6 percentage points (10.5 percent to 14.1 percent). Logistic regression confirmed significant main effects of group (OR = 2.6 for training vs. control, 95 percent CI 1.8 to 3.7, $p < 0.001$), campaign ($p < 0.001$), and a significant group-by-campaign interaction ($p = 0.003$), indicating that the reporting rate gap between groups widened significantly over time. Baseline reporting rates were indistinguishable between groups ($p = 0.72$), confirming that the observed divergence reflected the training intervention rather than pre-existing group differences.

Median time to report fell from 142 minutes at baseline to 47 minutes at the post-intervention campaign in the training group, a reduction of 95 minutes. The control group's median time to report was approximately stable across campaigns (148 to 136 minutes). Wilcoxon rank-sum tests confirmed that the training and control groups differed significantly in time to report at Week 4 ($p = 0.008$) and Week 8 ($p < 0.001$) but not at Week 1 ($p = 0.641$). The reduction in time to report in the training group is consistent with the video module's explicit emphasis on the organizational value of prompt reporting and with the behavioral decision framework that encouraged users to evaluate and report suspicious messages in a single workflow rather than deferring reporting.

C. Theme-Level Analysis of Training Effectiveness

To examine the theme-specificity of training effects more formally, CTR differences between training and control groups were calculated separately for each campaign's message theme. The generic delivery theme at Week 1 (baseline, pre-intervention) showed no significant group difference as expected. The authority theme at Week 4 showed a nonsignificant difference of 3.2 pp ($p = 0.171$), with

the training group's authority-message CTR remaining at 22.6 percent despite five weeks of elapsed time since the video module delivery. The scarcity theme at Week 8 showed a significant difference of 7.7 pp ($p = 0.002$). These theme-level results suggest that the training program was more effective at reducing susceptibility to scarcity-themed messages than to authority-themed messages within the 8-week study window, a pattern consistent with field evidence that authority cues engage a deeply socialized compliance heuristic that is more resistant to brief training interventions [8].

Among training group participants who clicked the Week 4 authority-themed message ($n = 69$, or 22.6 percent of the training group), 48 percent had also clicked the Week 1 baseline message and 52 percent had not clicked at baseline, suggesting that the authority-themed campaign was identifying a partially overlapping population of susceptible individuals rather than a distinct group not captured by the baseline campaign. This overlap pattern implies that authority-theme susceptibility is a relatively stable individual characteristic rather than a transient behavioral state, reinforcing the recommendation for role-specific, targeted interventions rather than relying on general awareness training to address authority-framed social engineering.

V. DISCUSSION

A. Training Effectiveness: Mechanism and Magnitude

The observed CTR reduction in the training group, from 28.3 percent to 14.7 percent across 8 weeks, and the control group's smaller decline from 27.1 percent to 22.4 percent, together imply a net training effect of approximately 7.7 percentage points for the scarcity-themed post-intervention campaign. This effect size is substantially larger than those reported in the large-scale healthcare organization experiment [2] and comparable to the effects documented in more focused, smaller-scale simulation studies. Several

features of the present study's design may account for this relative effectiveness. First, the combination of triggered embedded micro-training and a structured video module ensured that training content was both contextually timed and substantively comprehensive, addressing the two proposed explanations for the healthcare study's null finding: insufficient engagement time and insufficient content tailoring. Second, the explicit framing of reporting as a valued security behavior in the video module, combined with the absence of identifiable reporting of individual click data to management, appears to have successfully reduced stigma barriers to reporting, producing a near-tripling of reporting rates in the training group.

The theoretical mechanism underlying the effectiveness of triggered embedded training is consistent with the information processing account of phishing susceptibility [14]: the training intervention was delivered at the precise moment that a susceptibility-relevant behavior was performed, maximizing the motivational salience of the learning content and interrupting the habitual processing routine that produced the click. This mechanism predicts that training should be most effective for message types that engage habitual or heuristic processing, and least effective for those that trigger deeply ingrained social norms such as authority deference, a prediction confirmed by the theme-level results.

B. The Authority-Theme Challenge and Its Organizational Implications

The authority theme's relative resistance to the training intervention is the most practically significant finding of the study and one with direct implications for how organizations should allocate security investment. The nonsignificant 3.2 percentage point CTR reduction for authority-themed messages at Week 4, compared with the significant 7.7 percentage point reduction for scarcity-themed messages at Week 8, indicates that the brief training program used in this study was insufficient to override the authority compliance heuristic within the study's timeframe. This is consistent with Williams *et al.*'s [8] field finding that authority-framed spear-phishing messages consistently outperform other message types in generating clicks across diverse organizational populations.

The organizational implications are clear: authority-based social engineering should be treated as a distinct risk category requiring a dedicated response beyond what general awareness training can deliver. Role-specific simulation exercises that give employees practice in applying verification procedures when receiving requests that appear to come from senior leadership, combined with technical controls such as email authentication enforcement, external sender banners on inbound messages, and network-level blocking of spoofed executive domains via DNS filtering [4], represent a more appropriate response to authority-theme susceptibility than additional awareness training. These technical controls, combined with explicit organizational

policies on how senior leaders do and do not communicate urgent requests, can create environmental structures that reduce the behavioral load on individual employees in high-risk situations. The ethical dimensions of authority-based verification procedures, including the risk that excessive verification requirements create organizational friction that is itself exploited by social engineers, require the kind of principled professional judgment described in Blaise *et al.* [7].

C. Adversarial AI, IoT Risks, and the Future of Training Program Design

The 8-week study horizon and the use of conventional phishing message templates mean that the present findings may not generalize to the increasingly capable adversarial AI-generated phishing content that is emerging as the next generation of social engineering attack. Auwal's [3] systematic analysis of adversarial threat vectors identifies spear-phishing content generation as a high-impact, technically accessible capability in which large language models can produce messages that dynamically adapt to recipient-specific signals, including job title, recent online activity, organizational affiliations, and communication style patterns extracted from public sources. Training programs calibrated against known phishing templates may produce specific pattern-recognition skills that fail to generalize to novel, AI-generated content that does not match established phishing signatures. Future simulation campaigns should incorporate AI-generated message stimuli to test whether training-induced improvements in deliberate processing transfer to this new threat category or whether additional training content specific to AI-generated social engineering is required. Work on adversarial AI vulnerability in secondary education has likewise identified gaps in cybersecurity competencies and proposed frameworks that integrate threat recognition, behavioral resilience, and critical authentication [27].

The organizational security ecosystem also extends beyond email-based phishing. Auwal's [11] qualitative investigation of directory brute-forcing and artifact exposure documented how web application vulnerabilities create entry pathways that can be chained with phishing to escalate organizational compromise, and the companion SME vulnerability study [12] illustrated how resource-constrained organizations are particularly exposed to multi-vector attacks that combine phishing with exploitation of unpatched web application vulnerabilities. The research on smart medical IoT security further demonstrates that phishing-initiated compromise in connected healthcare environments can propagate to critical infrastructure through MITM attack vectors that are facilitated by weak endpoint security practices [17]. These findings collectively reinforce that organizational phishing training programs should be designed and evaluated as one component of a layered security strategy rather than as a standalone countermeasure. Organizational survey evidence on ransomware shows that data exfiltration is the strongest predictor of ransom payment [28], underscoring the cost of a

single successful click, and analysis of benchmark datasets for cyber attack prediction points to gaps that must be closed before AI-driven defenses can be deployed operationally [29].

D. Sociological and Governance Dimensions

The sociological framing of organizational phishing susceptibility developed by Auwal and Lazarus [5], [6] draws attention to dimensions of vulnerability that purely psychological accounts overlook. Employees who are under financial stress, who experience high workload pressure, or who work in organizational cultures that strongly reward deference to authority may be systematically more susceptible to phishing messages that exploit those specific vulnerabilities, regardless of their factual knowledge about phishing indicators. Training programs that address these contextual dimensions, for instance by normalizing verification behaviors and creating organizational norms that make it acceptable to question urgent requests regardless of apparent sender seniority, address a layer of susceptibility that cognitive and personality-focused training content cannot reach. The ideal victim literature [10] is relevant here: employees who have clicked a phishing link are often reluctant to report the incident because they fear being perceived as having failed a competence test rather than as a victim of deception. The reporting rate improvements observed in the training group suggest that deliberately reframing reporting as a valued, protective security behavior is an effective component of training design that deserves greater emphasis in program evaluation research.

The student and workforce perspective on AI-mediated training content [13] suggests that employees' perceptions of the fairness, transparency, and relevance of training programs significantly influence engagement and therefore training effectiveness. This is particularly relevant for simulation-based programs, which involve an element of deception that may be perceived negatively by employees who feel that their privacy or dignity has been violated. Organizational governance structures that make the rationale for simulated phishing campaigns transparent, that use click data only for program improvement rather than individual performance management, and that offer employees a genuine voice in how training programs are designed, are likely to produce greater employee engagement and therefore larger training effects than governance frameworks that treat simulations as covert performance evaluations.

E. Limitations

Several limitations qualify the interpretation and generalizability of the present findings. First, the study was conducted within a single organization of 612 employees. While this design is necessary for clean randomization and experimental control, it limits generalizability to larger, differently structured, or differently cultured organizations. The contrast with the healthcare study's null findings [2] illustrates how substantially effectiveness can vary across

organizational contexts, and future meta-analytic work comparing effect sizes across multiple field experiments is needed to characterize this heterogeneity. Second, the 8-week study duration captures only short-term effects of the training intervention.

Prior research on susceptibility change over time [15] suggests that training-induced behavior change can decay over subsequent months as the novelty of the intervention fades and as employees encounter real-world phishing messages that differ from simulation templates. Third, only one message exemplar per theme was used in each campaign, which limits the extent to which observed theme-level differences can be attributed to the theme category per se rather than to specific features of the message text or visual design. Fourth, the design did not measure individual-level psychological predictors such as impulsivity, which prior survey-based research has identified as a strong predictor of click rate [18], leaving open the question of whether training effectiveness interacts with individual psychological traits. Future work incorporating both experimental and individual-difference measures within the same study would enable a more complete account of when, for whom, and for which message types training is most effective.

VI. CONCLUSION

This paper reported an 8-week randomized field experiment evaluating a combined embedded micro-training and structured video module intervention in a mid-sized organization of 612 employees. The training group showed a statistically and operationally significant reduction in phishing click-through rate, from 28.3 percent to 14.7 percent, and a near-tripling of phishing reporting rate, from 9.8 percent to 31.2 percent, compared with smaller and mostly nonsignificant changes in the control group receiving standard policy reminders. Training effectiveness was theme-dependent: the generic delivery and scarcity-themed campaigns showed larger training-related improvements than the authority-themed campaign, reinforcing evidence that authority-based social engineering requires targeted supplementary interventions beyond what general awareness training can achieve.

These findings have several practical implications. Organizations should evaluate phishing training programs using theme-disaggregated outcome measures rather than relying on a single aggregate CTR figure that conceals meaningful variation in training effectiveness across cue types. Authority-based social engineering should be treated as a distinct risk category addressed through role-specific simulation exercises, explicit verification-procedure training, and technical controls including DNS and IP filtering [4].

The ethical governance of simulation campaigns must conform to principled professional frameworks [7] that maintain organizational trust, protect employee dignity, and ensure that simulation data are used for program improvement rather than punitive performance management.

Sustained, systematic cybersecurity education at the institutional level [19], [20] is a prerequisite for any simulation-based program to achieve lasting population-level behavior change. As adversarial AI tools continue to lower the barrier for producing highly convincing, recipient-tailored phishing content [3], future simulation programs will need to incorporate AI-generated stimuli to maintain ecological validity and to ensure that training-induced improvements in analytical processing transfer to the next generation of social engineering threats.

ACKNOWLEDGEMENT

The authors thank the participating organization's information security team for coordinating campaign deployment and providing anonymized click and reporting logs, and the university research ethics committee for guidance throughout the study design and conduct.

Declaration of Conflicting Interests

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Use of Artificial Intelligence (AI)-Assisted Technology for Manuscript Preparation

The author(s) utilised generative AI technologies solely to improve language clarity and refine grammar. No images were generated, manipulated, or altered using AI-assisted technologies.

REFERENCES

- [1] D. Hillman, Y. Harel, and E. Toch, "Evaluating organizational phishing awareness training on an enterprise scale," *Comput. Secur.*, vol. 132, p. 103364, 2023, doi: 10.1016/j.cose.2023.103364.
- [2] G. Ho, A. Mirian, E. Luo, et al., "Understanding the efficacy of phishing training in practice," in *Proc. IEEE Symp. Secur. Privacy (SP)*, 2025, pp. 76–95, doi: 10.1109/SP61157.2025.00076.
- [3] A. Muhammad Auwal, "Adversarial threat modeling in generative AI: A systematic mapping of attack vectors to defense mechanisms," *AI Ethics*, vol. 6, p. 284, 2026, doi: 10.1007/s43681-026-01154-8.
- [4] M. S. Islam, M. Sajjad, M. M. Hasan, and M. S. I. Mazumder, "Phishing attack detecting system using DNS and IP filtering," *Asian J. Comput. Sci. Technol.*, vol. 12, no. 1, pp. 16–20, 2023. [Online]. Available: <https://doi.org/10.51983/ajcst-2023.12.1.3552>.
- [5] A. M. Auwal and S. Lazarus, "Experiences of local victims of Yahoo Boys' socio-economic cybercrimes in Nigeria," *Discov. Psychol.*, vol. 5, art. 161, 2025, doi: 10.1007/s44202-025-00479-5.
- [6] M. Auwal and S. Lazarus, "Sociological and criminological research of victimization issues: Preliminary stage and new sphere of cybercrime categorization," *J. Digit. Technol. Law*, vol. 2, no. 4, pp. 915–942, 2024. [Online]. Available: <https://doi.org/10.21202/jdtl.2024.44>.
- [7] O. O. Blaise, I. Aaron, U. Alfred, and A. Amusa, "Evaluating the ethical frameworks of information security professionals: A comparative analysis," *Asian J. Comput. Sci. Technol.*, vol. 13, no. 2, pp. 61–66, Nov. 2024. [Online]. Available: <https://doi.org/10.70112/ajcst-2024.13.2.4289>.
- [8] E. J. Williams, J. Hinds, and A. N. Joinson, "Exploring susceptibility to phishing in the workplace," *Int. J. Hum. Comput. Stud.*, vol. 120, pp. 1–13, 2018, doi: 10.1016/j.ijhcs.2018.06.004.
- [9] L. E. Cohen and M. Felson, "Social change and crime rate trends: A routine activity approach," *Am. Sociol. Rev.*, vol. 44, no. 4, pp. 588–608, 1979.
- [10] N. Christie, "The ideal victim," in *From Crime Policy to Victim Policy*, E. A. Fattah, Ed. London: Palgrave Macmillan, 1986, pp. 17–30.
- [11] A. M. Auwal, "Directory brute-forcing and artifact exposure: Qualitative insight into underestimated threats to web application security," *Asian J. Eng. Appl. Technol.*, vol. 14, no. 1, pp. 28–32, 2025, doi: 10.70112/ajeat-2025.14.1.4296.
- [12] M. Auwal and J. Olaoluwa Abiodun, "A sectoral study of cyber threats and vulnerability management in resource-constrained Nigerian SMEs," *Asian J. Eng. Appl. Technol.*, vol. 14, no. 1, pp. 33–38, 2025, doi: 10.70112/ajeat-2025.14.1.4297.
- [13] M. Auwal, "Autonomy versus algorithm: A replication study of student perspectives on AI ethical boundaries," *Int. J. Educ. Technol. High. Educ.*, vol. 22, art. 72, 2025, doi: 10.1186/s41239-025-00570-w.
- [14] J. Wang, Y. Li, and H. R. Rao, "Testing individual differences in phishing vulnerability within an integrated information processing model," *Decis. Support Syst.*, vol. 51, no. 3, pp. 576–586, 2011, doi: 10.1016/j.dss.2011.03.002.
- [15] R. Chen, J. Gaia, and H. R. Rao, "An examination of the effect of recent phishing encounters on phishing susceptibility," *Decis. Support Syst.*, vol. 133, p. 113287, 2020, doi: 10.1016/j.dss.2020.113287.
- [16] F. Carroll, J. Adejobi, and R. Montasari, "How good are we at detecting a phishing attack? Investigating the evolving phishing attack email and why it continues to successfully deceive society," *SN Comput. Sci.*, vol. 3, art. 170, 2022, doi: 10.1007/s42979-022-01069-1.
- [17] A. M. Auwal, "Smart medical IoT security vulnerabilities: Real-time MITM attack analysis, lightweight encryption implementation, and practitioner perceptions in underdeveloped Nigerian healthcare systems," *arXiv preprint*, 2025, doi: 10.48550/arXiv.2510.09629.
- [18] Islam, M. M. Rashid, and F. Othman, "Identifying personality traits associated with phishing susceptibility," *Secur. J.*, vol. 38, art. 18, 2025, doi: 10.1057/s41284-025-00466-4.
- [19] K. A. Y. Yaseen, "Importance of cybersecurity in the higher education sector," *Asian J. Comput. Sci. Technol.*, vol. 11, no. 2, pp. 20–24, 2022. [Online]. Available: <https://doi.org/10.51983/ajcst-2022.11.2.3448>.
- [20] S. Ravichandran and K. L. N. Rao, "Design and development of an advancing web information stockpiling for engraved ontology in user contours," *Asian J. Comput. Sci. Technol.*, vol. 11, no. 2, pp. 11–15, 2022. [Online]. Available: <https://doi.org/10.51983/ajcst-2022.11.2.3379>.
- [21] A. M. Auwal and S. Shehu, "Measuring the impact of embedded phishing training on ransomware-lure click behavior: A field study across three organizations," *Asian J. Sci. Appl. Technol.*, vol. 15, no. 1, pp. 47–54, 2026, doi: 10.70112/ajsat-2026.15.1.4350.
- [22] A. M. Auwal and S. Shehu, "Psychological, situational, and demographic predictors of susceptibility to ransomware social-engineering lures: A survey study," *NeuroData*, vol. 3, 2026, doi: 10.63688/neurodata2026305.
- [23] A. M. Auwal, "Institutional typology of cybercrime enforcement in Nigeria using EFCC conviction records," *Discov. Data*, vol. 4, art. 19, 2026, doi: 10.1007/s44248-026-00115-0.
- [24] A. M. Auwal, "Invisible intrusions, internal threats, and ethical vulnerabilities in Nigeria's Internet of Things," *Discov. Internet Things*, vol. 5, art. 128, 2025, doi: 10.1007/s43926-025-00251-0.
- [25] A. M. Auwal, "IoT integration in telemedicine: Investigating remote patient monitoring and data transmission reliability," *Research Square preprint*, 2025, doi: 10.21203/rs.3.rs-3419693/v1.
- [26] A. M. Auwal, "A cross cultural extension study on privacy fatigue and AI adoption behavioral model," *Discov. Educ.*, 2026, doi: 10.1007/s44217-026-01903-1.
- [27] A. M. Auwal, "Adversarial AI vulnerability in secondary education: Threats and cybersecurity competencies," in *Secondary Education: Contemporary Issues, Practices, and Future Directions*. London, U.K.: IntechOpen, 2026, doi: 10.5772/intechopen.1016684.
- [28] A. M. Auwal and S. Shehu, "To pay or not to pay: An organizational survey of ransom payment decision drivers, data exfiltration, and the role of cyber insurance among mid sized organizations," *NeuroData*, vol. 3, 2026, doi: 10.63688/neurodata2026306.
- [29] A. M. Auwal, "Operationalizing cyber attack prediction: A gap-prioritized framework with dataset and model selection guidelines," *arXiv preprint arXiv:2606.03386*, 2026, doi: 10.48550/arXiv.2606.03386.