Phishing Email Awareness Systems

Posted on January 27, 2020

Introduction

Phishing email awareness systems are organizational controls designed to help people recognize, report, and recover from deceptive messages that attempt to steal credentials, deliver malware, redirect payments, or obtain sensitive information. The original essay correctly treats spear phishing as a serious social-engineering threat and proposes combining education with technical detection. It needs a more current and testable research design. Phishing is no longer limited to obviously fake email or a three-step sequence of message, website, and theft. Attackers may use legitimate cloud services, compromised accounts, business-email conversations, QR codes, telephone follow-up, multifactor-authentication fatigue, and artificial intelligence to make messages credible. No single algorithm can filter every attack without false positives. An effective awareness system therefore combines secure email architecture, identity controls, user-centered training, reporting tools, automated analysis, incident response, and metrics that account for message difficulty. The goal is resilience, not blaming an employee who encounters a carefully engineered deception. (Parsons et al., 2015)

Problem Definition

Organizations depend on email for approvals, invoices, document sharing, customer communication, password recovery, and executive decisions. This concentration of trust makes the inbox a valuable attack channel. A successful phishing message can provide an attacker with credentials or a foothold from which to move through systems, impersonate staff, alter payment instructions, or exfiltrate data. Traditional spam filters reduce volume but cannot interpret every organizational context. A message from a known supplier may be malicious because the supplier’s account was compromised. A legitimate-looking request may be suspicious only because it deviates from normal authority or timing. The research problem is therefore socio-technical: how can an organization reduce the probability and impact of phishing while preserving usable communication?

Research Questions

The project can be organized around five questions. First, which technical and contextual features most reliably distinguish suspicious messages in the target organization? Second, how should automated scoring, secure-email controls, and human reporting be combined? Third, how can awareness exercises reflect realistic difficulty without humiliating employees? Fourth, which metrics demonstrate risk reduction more accurately than click rate alone? Fifth, how should the system

respond after a user opens a link, submits credentials, or reports a message? These questions preserve the original thesis purpose while moving beyond the unsupported promise of one “specially crafted algorithm” that solves phishing universally.

Phishing, Spear Phishing, and Business Email Compromise

Phishing is a broad category of deceptive communication intended to induce an unsafe action. Spear phishing targets a selected person or group using relevant details. Whaling focuses on senior or financially powerful individuals. Business email compromise commonly involves impersonating an executive, supplier, or employee to redirect funds or data. Credential phishing sends users to a fake or adversary-controlled login. Malware phishing uses attachments or links to install code. These categories overlap. Classification matters because defenses differ: attachment sandboxing may reduce malware, while payment verification and separation of duties are more important for invoice fraud.

The Attack Lifecycle

A phishing campaign begins with reconnaissance and preparation. Attackers identify roles, relationships, language, and technology through public websites, social media, breaches, or compromised accounts. They build or acquire infrastructure, craft a pretext, and deliver the message. The target interprets cues and decides whether to click, reply, open, scan, pay, or report. If the action succeeds, the attacker may capture credentials, establish persistence, exploit access, and pursue an objective. Defenders can intervene at every stage. Domain protection, filtering, browser isolation, multifactor authentication, rapid reporting, session revocation, and transaction verification create layers so that one mistaken action does not automatically become a major incident. (National Institute of Standards and Technology, 2022)

Human Detection Cues

Common warning signs include unexpected urgency, pressure to bypass procedure, unusual sender domains, requests for secrets, mismatched links, unfamiliar attachments, altered payment instructions, and conversation that does not fit the sender’s normal behavior. These cues are useful but not universal. Attackers can use correct grammar, genuine branding, and compromised accounts. Training that presents phishing as a puzzle of spelling errors prepares users mainly for easy examples. Employees should instead learn to pause when a request changes authority, money, access, or secrecy. Verification should occur through a trusted channel obtained independently of the message.

The NIST Phish Scale

NIST developed the Phish Scale to rate the human difficulty of simulated phishing messages. The method considers the number and quality of observable cues together with how closely the message aligns with the recipient's work context. This is important because a low click rate on an obvious simulation does not prove strong awareness, while a high click rate on a highly convincing message does not prove that staff learned nothing. Applying a difficulty rating allows program managers to compare exercises more fairly and design a progression from recognizable threats to realistic contextual attacks. It also reduces the temptation to use deceptive "gotcha" exercises merely to create dramatic results. (Dawkins & Jacobs, 2023)

Secure Email Gateway

A secure email gateway can evaluate sender reputation, authentication results, links, attachments, language patterns, and known threat intelligence. Attachments may be scanned or detonated in controlled environments, while URLs can be rewritten for time-of-click inspection. These controls reduce exposure but create false positives and cannot know every legitimate business relationship. High-risk messages can be quarantined or marked with context, but warning banners should be designed carefully. If every external message displays the same warning, employees learn to ignore it. The gateway should generate explainable signals and support analyst review of unusual patterns.

Email Authentication

Sender Policy Framework, DomainKeys Identified Mail, and Domain-based Message Authentication, Reporting and Conformance help receiving systems evaluate whether messages using a domain are authorized and whether protected content changed in transit. DMARC also provides policy and reporting mechanisms. These standards make certain forms of direct domain spoofing more difficult, but they do not prevent look-alike domains or abuse of a legitimate compromised account. Organizations should protect their own domains, monitor reports, and manage third-party senders before enforcing stricter policy. Email authentication is a foundation rather than a complete anti-phishing solution.

Identity and Access Controls

Strong authentication limits the value of stolen passwords. Phishing-resistant methods, such as properly implemented security keys or passkeys, provide greater protection than codes that users can be tricked into entering on an attacker's site. Multifactor authentication should be combined with conditional access, device trust, session monitoring, and rapid revocation. Attackers may attempt repeated prompts or steal session tokens, so organizations cannot assume that any form of MFA

makes phishing harmless. Privileged accounts need additional controls, and access should follow least privilege so that one compromised identity cannot reach every resource.

Reporting Interface

Employees need a simple way to report suspicious messages from the email client or mobile device. The system should preserve useful headers and content, notify the reporter that the submission was received, and route the message for automated and human analysis. Feedback matters. If users never learn whether reports were useful, participation declines. The interface should also support urgent reporting after an unsafe action. A button labeled only “report phishing” may confuse users who are uncertain; guidance can explain that it is appropriate to report a message for review even when they are not sure.

Automated Triage

A triage engine can extract indicators, compare URLs and attachments with intelligence, check authentication, identify similar messages, and estimate risk. Machine learning may assist classification, but its performance depends on training data, drift, adversarial adaptation, and organizational context. The system should not silently delete high-impact messages based only on an opaque score. It can prioritize analyst attention and automate low-risk actions under defined rules. Logging, version control, privacy review, and testing are required. The research should compare the proposed model with baseline filters and report precision, recall, false-positive rate, false-negative rate, and time to decision.

Natural-Language and Behavioral Analysis

Text analysis can identify urgency, credential requests, financial language, impersonation, and unusual communication patterns. Behavioral models can compare a message with established sender-recipient relationships or payment workflows. These methods introduce privacy and governance concerns because routine employee communication may be analyzed. Collection should be minimized, access controlled, and use limited to cybersecurity. A behavioral anomaly is not proof of malicious intent; new projects and legitimate emergencies can also look unusual. Analysts need context and a way to correct the model.

Awareness Training Design

Annual slide presentations are insufficient for a changing threat. Effective training is brief, role-relevant, repeated, accessible, and connected with real reporting procedures. Finance staff need invoice and payment scenarios, while administrators may face document-sharing lures and executives

may receive legal or acquisition pretexts. Training should explain why attackers' techniques work and give users actions they can perform under pressure. It should not teach that every external link is forbidden, because modern work requires links. The objective is informed verification and rapid reporting.

Simulated Phishing Exercises

Simulations can provide practice and reveal process weaknesses, but they should be governed ethically. Campaigns should avoid exploiting personal trauma, medical fear, or employment insecurity merely to increase clicks. Results should be used primarily for improvement rather than public ranking or punishment. Individuals who interact with a simulation can receive immediate, concise coaching. Program managers should vary difficulty, channel, role, and technique. They should also test whether staff report the message and whether security teams can respond, not only whether someone clicked.

Metrics Beyond Click Rate

Click rate is easy to calculate but incomplete. A program should measure reporting rate, time to first report, time to remove similar messages, credential submission, repeated susceptibility, detection difficulty, affected roles, and incident consequences. Technical measures include authentication coverage, gateway detection, false positives, MFA resistance, and compromised-account containment time. Qualitative feedback can reveal why a message was convincing and whether staff understand the process. A decrease in clicks may reflect easier simulations rather than improved ability, which is why the Phish Scale and contextual measures are valuable.

Incident Response After Interaction

When a user reports that they clicked or entered credentials, the organization should respond without delay or humiliation. Actions may include resetting or disabling credentials, revoking sessions, reviewing mailbox rules, isolating devices, searching for related messages, preserving evidence, blocking infrastructure, notifying affected parties, and monitoring transactions. The user's prompt disclosure is a protective behavior. Punitive reactions encourage concealment and allow attackers more time. Response plans should distinguish a simple visit to a harmless page from credential theft, malware execution, financial fraud, or data disclosure.

Business Process Controls

Technical email security cannot prevent every fraudulent request. High-risk transactions should require controls outside the email itself. Changes to bank accounts can be verified through a known

telephone number, large payments can require separation of duties, and sensitive-data requests can follow approval workflows. Executives should not be able to override controls casually through urgent messages. Attackers exploit organizational culture, especially fear of authority and reluctance to question senior staff. A secure culture explicitly authorizes employees to verify unusual requests.

Privacy and Ethical Considerations

Phishing systems may inspect communications, track simulation behavior, and create risk profiles. Employees should know the purpose, data collected, retention, access, and how results are used. Monitoring should not expand quietly into productivity surveillance. Simulations should accommodate disability and language differences and should not stigmatize older staff or nontechnical roles. Security responsibility is shared: designers, administrators, leaders, vendors, and users all influence risk. The phrase “human vulnerability” should not become a justification for neglecting technical and organizational controls.

Research Method

A mixed-method evaluation would be stronger than a literature review alone. The project can begin with analysis of anonymized incident and simulation data, interviews with security staff and employees, and review of email architecture. A prototype can then be tested in a controlled environment against a labeled dataset and compared with existing filtering. A phased pilot can measure reporting, response time, false positives, and user experience. Interviews after exercises can explain why users made decisions. Institutional approval, privacy safeguards, and secure data handling are necessary because messages may contain confidential information. (“Cybersecurity and Infrastructure Security”, 2025)

Proposed System Architecture

The proposed system should include domain authentication, a secure gateway, time-of-click URL analysis, attachment inspection, a reporting button, automated triage, identity controls, analyst workflow, and incident-response integration. A management interface can show trends and allow approved changes to rules, but simple blacklists and whitelists should not be the primary strategy because compromised trusted domains are common. The system should maintain confidence scores, evidence, and expiration for indicators. High-impact actions require review, while rapid containment can be automated under tested conditions. Architecture should be modular so that one vendor or detection model does not become a single point of failure.

Limitations

Phishing evolves in response to defense, so findings from one organization or period may not generalize. Simulations cannot reproduce all emotional and operational pressures of real attacks. Employees who know a program exists may behave differently, and click data do not measure every risky action. Machine-learning performance can decline as language and infrastructure change. Privacy constraints may limit behavioral analysis. These limitations should be reported rather than hidden. The objective is not a claim of perfect prevention but measurable reduction in likelihood, detection time, and consequence.

Conclusion

Phishing email awareness systems are most effective when they combine human judgment with layered technical and organizational controls. Modern attacks use compromised accounts, cloud services, payment pretexts, QR codes, and artificial intelligence, making simple lists of spelling errors obsolete. A resilient program protects domains, filters messages, uses phishing-resistant authentication, gives employees an easy reporting channel, and responds rapidly after interaction. Training should be realistic, ethical, and evaluated according to message difficulty rather than click rate alone. The proposed research should test an integrated architecture through technical performance, user behavior, incident response, and privacy analysis. No algorithm can identify every malicious message, but a well-designed system can ensure that one deceptive email does not become an uncontrolled organizational breach.

References

Cybersecurity and Infrastructure Security Agency. (2025). *Phishing guidance and Secure Our World resources*.

Dawkins, S., & Jacobs, J. (2023). *NIST Phish Scale user guide* (NIST Technical Note 2276). National Institute of Standards and Technology.

National Institute of Standards and Technology. (2022). *Federal cybersecurity awareness programs: A mixed methods research study* (NISTIR 8420).

Parsons, K., McCormac, A., Pattinson, M., Butavicius, M., & Jerram, C. (2015). The design of phishing studies. *Human-Centric Computing and Information Sciences, 5*, 1–17.

Verizon. (2025). *Data breach investigations report*.