Judging Phishing Under Uncertainty: How Do Users Handle Inaccurate Automated Advice?

Tarini Saka
University of Edinburgh
Edinburgh, United Kingdom
tarini.saka@ed.ac.uk

Kalliopi Vakali
University of Edinburgh
Edinburgh, United Kingdom
K.Vakali@sms.ed.ac.uk

Adam Jenkins
King's College London
London, United Kingdom
adam.jenkins@kcl.ac.uk

Nadin Kokciyan
University of Edinburgh
Edinburgh, United Kingdom
nadin.kokciyan@ed.ac.uk

Kami Vaniea
University of Waterloo
Waterloo, Canada
kami.vaniea@uwaterloo.ca

ABSTRACT

Providing accurate and actionable advice about phishing emails is challenging. The majority of advice is generic and hard to implement. Phishing emails that pass through filters and land in user inboxes are usually sophisticated and exploit differences between how humans and computers interpret emails. Therefore, users need accurate and relevant guidance to take the right action. This study investigates the effectiveness of guidance based on features extracted from emails, which even in AI-driven systems can sometimes be inaccurate, leading to poor advice. We examined three conditions: control (generic advice), perfect advice, and realistic advice, through an online survey of 489 participants on Prolific, and measured user accuracy and confidence in phishing detection with and without guidance. Our findings indicate that having advice specific to the email is more effective than generic guidance (control). Inaccuracies in the guidance can also impact user decisions and reduce detection accuracy.

CCS CONCEPTS

• **Security and privacy → Phishing**; *Usability in security and privacy*; • **Human-centered computing → User studies**.

KEYWORDS

Phishing; User Guidance; Security, Attack Detection

ACM Reference Format:
Tarini Saka, Kalliopi Vakali, Adam Jenkins, Nadin Kokciyan, and Kami Vaniea. 2025. Judging Phishing Under Uncertainty: How Do Users Handle Inaccurate Automated Advice?. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3706598.3714267

1 INTRODUCTION

Phishing is a cyber-attack where criminals send deceptive messages to individuals, posing as legitimate sources, in order to obtain sensitive information or distribute malware, while the majority of such messages are automatically detected and deleted, some reach users who must decide if they are phishing or benign. Email is the most common method used for phishing attacks [84], making it crucial to provide users with guidance on identifying phishing emails. Existing email guidance systems are often either generic or automated. Generic systems provide the same guidance regardless of the specific email [5, 72, 92], while automated systems use methods such as blacklists of phishing domains or AI-based analysis [12, 44, 62, 64, 87]. Past research has focused on the impact of education [57], browser-based warnings [3, 28, 94], and URL warnings [6, 64], with limited exploration of automated email analysis for guidance. There has been little examination of the impact of inaccurate information in guidance systems on human decision-making. This research addresses this gap through an online experiment comparing a hypothetical `Perfect` email analysis report to a generic `Control` guidance and a `Realistic` report containing occasional inaccuracies and errors.

According to Egress's 2024 Email Security Risk Report [32], "96% of surveyed organizations experienced negative impacts from phishing attacks." Additionally, 58% of organizations fell victim to account takeover attacks, with 79% of these originating from a phishing email and 83% bypassing multi-factor authentication. Hence, phishing is a major concern for organizations. To combat these attacks, it's crucial for organizations to implement multi-layered security measures [20, 60]. In other words, one solution is not enough. It is recommended to use a set of solutions including automatic detection, training, and a robust phishing report workflow that enables fast reaction to phishing attacks.

Automated phishing detection and filtering is very effective though not perfect. Prior research has developed advanced algorithms, rule-based filters, and blocked domain lists [8, 35, 46]; however, phishing emails still manage to evade defences and end up in users' inboxes. Attackers constantly change their tactics creating increasingly sophisticated and personalized phishing attacks. As a result, the final responsibility for phishing identification rests with end-users, and their judgements (e.g. phish or benign) can have significant consequences for other users and their organization [36, 39, 65]. Educating users to recognise and report phishing emails is essential for both personal protection and organizational security. Considerable research has been dedicated to creating tools to assist users in this critical activity, including security and phishing-specific training tools that gamify decision-making [19, 76], as well as contextual warnings to inform the

CHI '25, April 26-May 1, 2025, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3714267

user [41, 44]. In current systems, if users need help assessing an email, they must report it and ask for advice, which is not ideal. First, the reporting process often takes time or may not yield a response at all [5]. Second, research indicates that users tend to not report phishing due to a lack of confidence in their ability to identify legitimate phish [50]. Third, many organizations utilize an auto-reply system containing lengthy and generic information about phishing, along with standard advice and a promise of follow-up which may never happen. This approach may not inspire confidence in the system as users receive the same response for every email.

A possible solution is to have a system that can provide in-the-moment contextual guidance or advice to users. Recent advancements in artificial intelligence (AI), especially in machine learning and natural language processing, have created the opportunity to develop automated guidance tools that can assist users in leveraging their contextual information effectively. AI provides an opportunity to replicate some of the approaches human experts use to identify phishing and therefore create more tailored guidance and recommendations for users [44, 62, 69]. These works have focused on the technical aspects of such systems like which computer-focused features are needed [69, 79] or methodology approaches like LLMs [44, 62, 71]. How to best inform the user of the outcome of such technical analysis in a way that supports their decision-making process has been less well studied [42] though interesting user interface ideas have been proposed [41].

A significant challenge when implementing AI-based systems for phishing guidance is the potential impact of uncertainty in AI-generated suggestions or predictions. Especially in a space where the attacker is likely designing communications with the goal of deceiving AI filters. Current approaches often rely on disclaimers such as, "ChatGPT can make mistakes. Check important info,"[1] which serve more as legal safeguards than meaningful guidance for users. Perceived inaccuracies can also undermine user confidence in the system [22, 27, 47]. A key contribution of this work is the exploration of the impact of showing users AI-drawn conclusions which have a probability attached to them, along with the key features used to draw that conclusion. The findings from this study provide crucial insights that should guide the technical development of these tools, ensuring they are both effective and user-friendly. Additionally, this work offers valuable information for organizations and security platforms considering the adoption of such tools, helping them understand the potential benefits and drawbacks.

In this study, we evaluate the impact of an automated analysis report on an individual's ability to detect phishing emails, as well as their confidence. We focus on the benefits of tailored, email-specific analysis compared to the commonly used generic guidance, and we also examine how incorrect conclusions in the automated analysis influence user decision-making. We conducted an online experiment with 489 users to examine how their phishing identification and confidence changes with the introduction of 1) generic guidance (`Control`), 2) automatic accurate feedback (`Perfect`), and 3) automatic feedback with some incorrect conclusions (`Realistic`). Our research questions are:

RQ1: To what extent does having on-demand in-the-moment phishing guidance impact users' accuracy and confidence in identifying phishing emails?

RQ2: To what extent does tailored email analysis influence users' accuracy and confidence in distinguishing phishing from benign emails, compared to generic advice?

RQ3: To what extent does the accuracy of tailored email analysis and presented facts impact the user's assessment of an email being phishing or benign?

We find that providing accurate and contextual guidance helps participants better identify whether emails are phishing or benign. Guidance with some inaccuracies still led to some improvement compared to providing no guidance at all. Initially, users had more trouble identifying benign emails correctly when they had no guidance, but once they had access to contextual guidance, their accuracy in identifying benign emails improved. In general, access to contextual guidance helped improve both accuracy and confidence. However, some inaccuracies in the guidance led to reduced accuracy and confidence. Offering users contextual reports seems promising for increasing user accuracy in identifying potential phishing attempts as well as boosting their confidence in deciding on the safety of an email.

2 BACKGROUND AND RELATED WORK

In this section, we begin by addressing the importance of phishing reporting and the crucial role users and training play in this process. We will then explore various features commonly used for user guidance, the differences in perspectives between expert and non-expert users, and the necessity for automated email analysis systems.

2.1 Phishing Reporting

Phishing reporting is the process through which users identify and flag suspicious emails to alert the appropriate organizations or authorities [5, 51]. This practice is essential in cybersecurity, as it enables IT staff to quickly recognize ongoing attacks, mitigate threats in real-time, and block compromised accounts and incoming emails [14, 15, 24, 26, 45, 50, 78]. Researchers have sought to understand the factors that lead users to report [24, 45, 56], as well as those that discourage them from reporting incidents [26]. Although some users report emails just to verify their legitimacy [14, 15], reporting rates are still suboptimal. This is largely due to fears of the consequences of misreporting, lack of trust in IT [50], and unclear reporting mechanisms [50, 78]. One way to improve reporting rates is to train users to recognize phishing emails [10, 21].

2.2 AI-Assisted Phishing Detection

Phishing is particularly challenging because attackers constantly adjust their tactics to bypass security filters, ensuring that malicious emails often end up in users' inboxes. They often design emails that present one context to the automated filters and a different one to the recipient. This puts the responsibility on users to make the final decision on whether the email is legitimate or malicious, a task that carries serious consequences if mishandled, such as data breaches or financial loss. However, not all users are good at identifying phishing and often require guidance to make the right decision.

[1] https://chatgpt.com/ Accessed Dec 4, 2024.

An email in an inbox has passed through security filters and likely checked for authentication fails, against blacklisted domains, and passed these checks. However, even the most advanced phishing emails that manage to evade filters and checklists are often caught by experienced security experts. These experts utilize a combination of technical and contextual indicators to identify and make a decision [91, 93]. For example, consider an email that appears to come from a well-known brand. While a person might notice that the email's sender domain and URLs do not align with the brand's official site, a computer can analyze the text and the sender, without drawing this conclusion. If an AI system could replicate this thought process and present such contextual details, it would be much easier to guide users toward the correct decision [69, 71]. Such AI-assisted decision-making would combine the individual strengths of humans and the AI to optimize the outcome [96] and could be very crucial to improving phishing reporting rates.

2.3 Real-time Tailored User Assistance

A significant factor contributing to the success of phishing attacks is the inability of users to identify phishing emails, either due to a lack of attentiveness or expertise. The issue of inattention is usually addressed through the use of security warnings, commonly in the form of browser-based or banner warnings. Browser warnings typically appear after a user has already clicked on a link, which may sometimes be too late [3, 28, 34, 95]. On the other hand, banner warnings alert users to the potentially suspicious nature of an email as soon as it is opened. However, these warnings often fail to explain specific suspicious elements, placing additional responsibility on the user. Petelka *et al.* [64] found that warnings focused on links reduce phishing click-through rates compared to banner warnings, with forced attention warnings being the most effective. However, as organizations receive emails from various domains and phishers increasingly employ URL obfuscation techniques such as redirection [6] and shortening [52, 55], relying solely on URLs is insufficient [77].

To address the lack of knowledge problem, user training is the recommended approach [38, 58]. Training delivery methods can be classified into persistent (or embedded training) and standalone methods. Standalone methods can be further defined by the degree of user engagement, such as passive (written training materials and educational videos) or interactive (e-learning and educational games) [43, 58]. However, the effectiveness of training on its own tends to diminish after a few months, as indicated by various studies [68]. Employee training is often provided up-front [67] or after an employee falls for mock phishing [49] and is not always effective, especially for non-expert users with little technical experience. Parsons *et al.* [63] studied the key cues that users utilize to differentiate between phishing and genuine emails and found that participants often use poor indicators of legitimacy, highlighting the need for targeted education and training in recognizing phishing threats. In their proposed Phishing Susceptibility Model, Zhuo *et al.* [97] define three temporal stages that explain human vulnerability to phishing attacks. One of their key findings is the research gap around the effectiveness of in-the-moment assistant tools. Another important aspect of an efficient guidance tool is to create well-tailored contextual advice [29, 36, 41, 74]. Franz *et al.* [36] highlighted a gap in the literature regarding tailored user interventions for preventing phishing attacks. They advocate that using tailored advice instead of one-size-fits-all interventions will enhance such systems. To address this issue, Jenkins *et al.* [41] propose the idea for a phishing-advice tool, PhishEd, to provide quick and accurate support to those who report phishing attempts. However, their poster does not include any implementation or evaluation details. Kashapov et al. [44] utilized transformer-based machine learning models to examine potential psychological triggers, identify potential malicious intent, and generate concise summaries of emails. Their goal was to help users determine if an email is suspicious and also learn about more advanced malicious patterns. The study yielded promising results, and the researchers suggest further investigation through user studies and objective experimental analysis [44]. Jayatilaka *et al.* suggest that anti-phishing education, training, and awareness should be tailored to address the diversity and complexity of how individuals respond to phishing attempts. Their findings indicate that people exhibit varied difficulties: some struggle to identify the legitimacy of emails, others have trouble validating emails, and some fail to take safe actions even after correctly judging an email's legitimacy. Therefore, a one-size-fits-all approach is insufficient [40].

2.4 AI Systems and Inaccuracies

Inaccuracies are inevitable in any automated system [53], and an automated email guidance system is no exception. The accuracy and results can vary based on factors such as the training data, model selection, and feature set [18, 35, 46]. Our focus was primarily on two types of errors: incorrectly parsed information and misclassification of phishing or benign emails. The former is when details such as the organization, URL, or organization name may be misinterpreted [11]. The second error occurs in the classification of the email, where the system assigns a likelihood of the email being safe or malicious [1, 35, 46]. While a false positive (classifying a safe email as phishing) may lead to minor inconveniences, a false negative (classifying a phishing email as safe) poses serious risks. Despite the prevalence of automated systems that offer banner warnings, browser alerts, or URL advisories, previous research has not extensively explored the effects of incorrect classifications on user behaviour and decision-making. Our contribution lies in evaluating not only the effectiveness of the guidance but also the feedback from users when exposed to errors, providing insight into the long-term viability of such systems.

For a user guidance system to be effective over time, it is crucial that users have confidence in the system. This confidence is directly influenced by the quality and accuracy of the information provided. If the guidance frequently contains errors or inconsistencies, users will likely lose trust in its reliability, undermining its effectiveness [2, 22]. Rechkemmer and Yin [66] found that an individual's belief in a model's predictions is significantly influenced by the model's confidence in the predictions and that the model's actual performance metrics, such as accuracy, have a strong impact on how often people follow the model and trust it overall. Furthermore, Zhang *et al.* [96] found that providing confidence scores helps users calibrate their trust in an AI model. They argue that the improvement of AI-assisted decision-making also depends

on whether humans can contribute unique knowledge to complement the AI's errors. In our study, we explore these implications by incorporating several key elements in the guidance report, such as likelihood percentages of scam classification and supporting evidence for the classification. These features were designed to help users evaluate the credibility of an email, and apply their external knowledge and instincts to make their final decision.

3 REPORT STRUCTURE

A key goal of this study is to explore how humans can be assisted with automated solutions such as AI-driven technologies. Phishing is a good possible candidate for such AI-assisted decision-making [80, 96] because there are some factual information in emails that computers are better able to reason about (e.g., inspecting the sender domain) while some other information could be captured by humans easily (e.g., the visual cues in an email). The report design is based on an effort to support this collaboration.

Phishing emails have multiple characteristics or features, some of which are readily apparent to users, while others are more subtle or hidden. Visible features include elements like the "from" address, while subtler features include the destination of embedded links, which can be harder to discern. Additionally, some aspects, such as DMARC and DKIM cryptographic signatures, are effectively invisible to the average user. Similarly, there is contextual information in the emails available to the user that the computer is unaware such as what bank they use, what a normal email looks like from their organization, or if they are expecting a package. Correctly judging if an email is phishing or not requires knowledge about a range of features, contextual information, and possible scams the email could be. Phishing emails are sometimes deliberately crafted to appear differently to human users and automated systems. For example, Figure 1 shows an email designed to mislead both parties: it displays as a Starbucks email to the user but includes white-on-white text meant to deceive email filters into thinking it is about Greek food. Computers struggle to identify all phishing attempts partly because they lack some of the relevant information that is easily accessible to users. For instance, in Figure 1, a user might perceive the email as originating from a legitimate Starbucks-contracted survey but a computer will fail to understand what the user perceives.

3.1 Report Concept

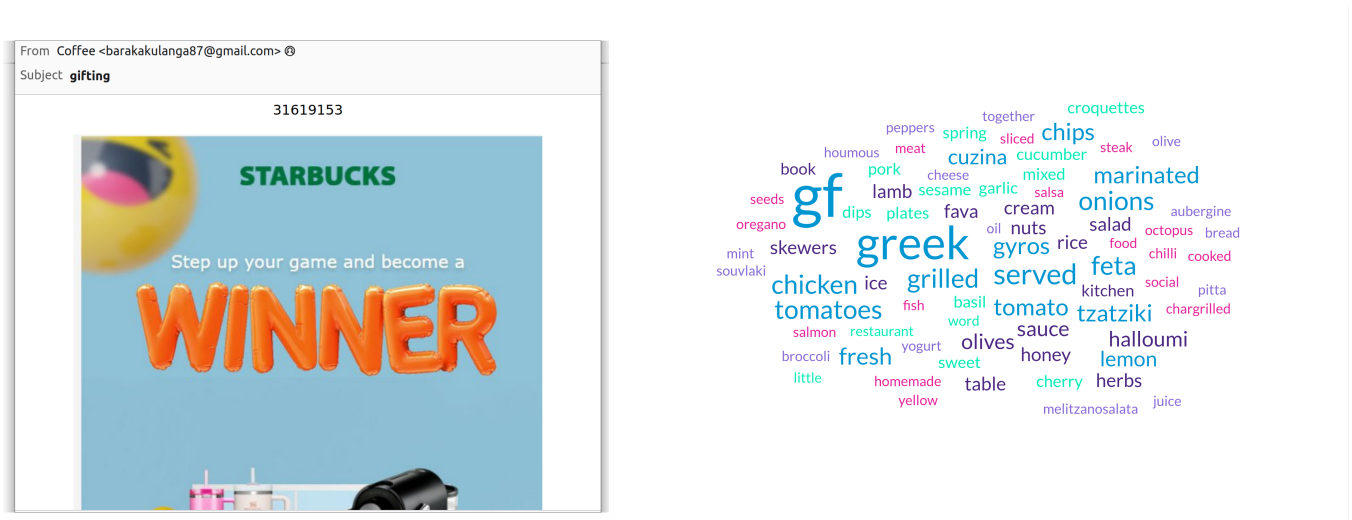
The tested report is based on the concept of human-AI collaboration, which leverages the complementary strengths of both humans and artificial intelligence to tackle tasks that might be difficult for either to handle independently. The idea is that when a user encounters a potentially suspicious message and is uncertain about its safety. Instead of making a decision based solely on their judgment or curiosity [88], the user can request the system to generate a report to help them decide. The AI system then automatically analyses the email, detects specific cues about the email and produces a recommendation based on those facts. Such facts would be based on contextual features that emulate expert analysis rather than the usual technical indicators based on blacklists and authentication signatures. Ideally, the AI's capabilities would be 100% accurate, ensuring that every recommendation is completely reliable. However, given the adversarial nature of phishing, there will be cases of inaccuracy. However, due to the ever-evolving adversarial nature of phishing attacks and the probabilistic nature of many AI models, achieving perfect accuracy is challenging.

3.2 Report Design

The report used in the study (Figure 2b) is a simple implementation of the idea of showing users information about a phishing email and advice based on that information. It features 4 boxes which contain automatically extracted information about the potential phishing email and a large box at the bottom that advises if the email is likely safe or a scam. If it is likely a scam, then the type of scam is explained. We chose to provide users with an email classification likelihood, along with extracted information to support this classification, in order to enhance their understanding of the report. Wang *et al.* [90] examined AI-assisted decision-making systems and identified three key properties that effective explanations should satisfy: improving understanding of the AI model, recognizing model uncertainty, and fostering calibrated trust. Their findings highlight that the effectiveness of explanations varies with users' domain expertise, implying that for a general audience, suggesting that explanations for a general audience should prioritize simplicity and clarity.

The report starts with a statement that it is automatically generated to make it clear that the information should be treated as if a computer generated it without human review. Next, four boxes show: who the email claims to be from, the action(s) the user is being asked to take, the domain of the from address, and the destination of links in the email. The information was selected to represent the information used by most expert and non-expert users to scrutinize an email [65, 91, 93]. In one case, we used the subject line as evidence rather than the sender domain, which was spoofed, as it was stronger evidence for its classification. The sender domain and link destination information also provide comparison information to highlight expectation violations. For instance, in Figure 2b the email claims to be from the *<Blinded>* University but the sender domain is not a University domain. Finally, the report provides a *high-level classification of the email* as 'Safe' or a likely 'Scam' along with a confidence score. This part of the report provides a probabilistic assessment of the email's safety based on the extracted information. If the email is detected as a scam, it also provides a short explanation of the most likely scam, as these details can improve confidence in the system and also teach the user for future instances. In case the email is classified as 'Safe', we reassure the user that the email is safe to interact with, while also providing a resource for further assistance if they remain uncertain.

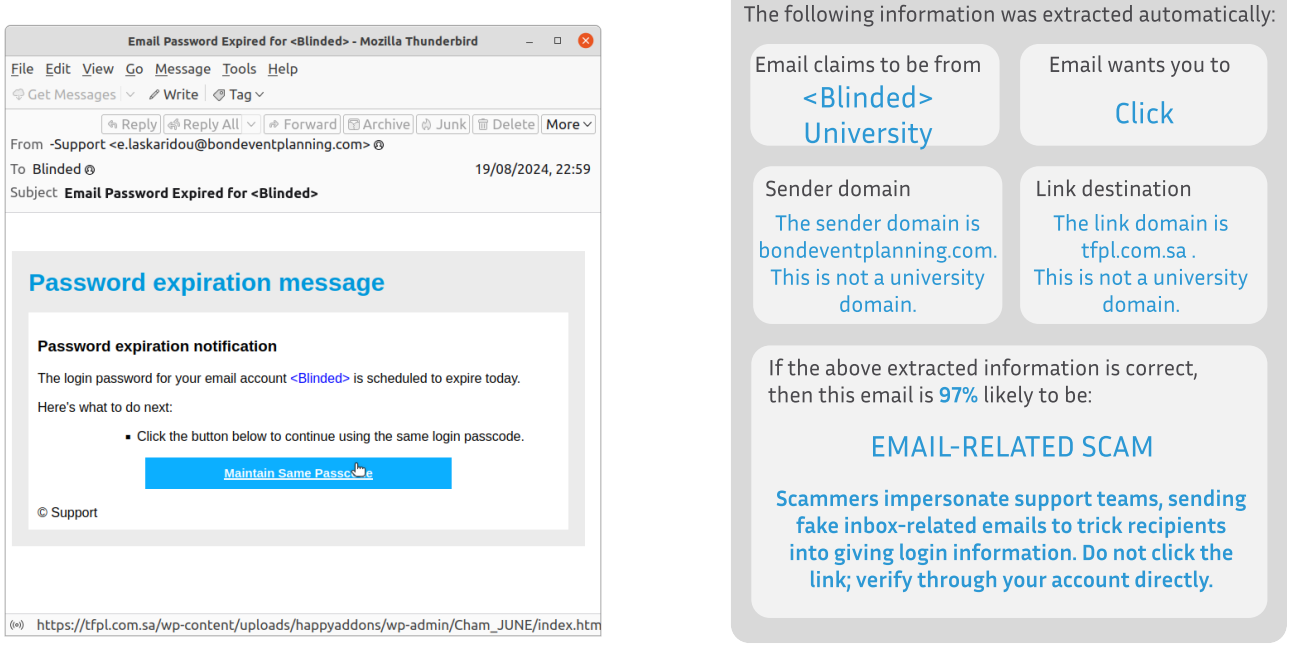
The concept behind the automated report is that AI can be used to automatically detect critical information in an email, such as who the email claims to be from, what it wants the user to do, sender information, and link domains. For a benign email, there is typically internal consistency between such information. A legitimate email from a bank will have a bank domain in the sender's address and will direct links to official bank URLs. In contrast, Figure 1 illustrates a scenario where such consistency is not present. The report's structure was designed based on key features of emails that could potentially be automatically extracted from an email but may have some degree of error. For example, extracting the sender address from an email is easy. The email in Figure 2a claims to be from the IT



(a) Phishing email impersonating Starbucks using an image.

(b) Wordcloud of the email text.

Figure 1: An example of a phishing email where an image is used to convey one meaning to the human and white-on-white text conveys a different meaning to the computer. The use of a Gmail from address is a clear indication that this email is not from Starbucks.



(a) A sample phishing email from the dataset used.

(b) `Perfect` report - Tailored to the specific email.

Figure 2: The two types of reports used in the experiments. (1) `Control` report consisting of generic advice (2) `Perfect` condition report, tailored to the specific email. The `Realistic` condition report has the same template with a different analysis.

support team at *<Blinded>* University, but that information is not obvious from the email address *e.laskaridou@bondeventplanning.com* alone.

The selection of features included in our hypothetical report was based on a combination of experience, insights from interviews with IT staff, and existing literature in the field [62, 69, 79, 86]. Two common questions users are encouraged to consider when they suspect an email is who it is from (sender domain and organization name) and if they were expecting it (the main topic of the email) [79, 91]. Our choice of features was influenced by this practical guidance and supported by related work in the literature.

To ensure that such a report is technically possible to create, a master's student conducted a prior study that explored automatically extracting phishing indicators from an email and presenting them to users in a browser-based email interface [81]. They found indicators similar to the ones in this study to be possible to extract automatically. They also conducted a pilot study with 22 participants to see if their report would help people judge emails and found that users showed increased precision and confidence in detecting phishing attempts, suggesting the potential of automated user assistance to reduce human error in email security [82].

4 STUDY DESIGN

We conducted an online survey using the Prolific platform[2]. The experiment examined how different types of guidance reports (`Control`, `Perfect` report, and `Realistic` report) influenced participants' ability to determine whether an email was phishing or not. Initially, all participants reviewed the same set of 10 emails, 4 of which were phishing, without any assistance. Following this first round, participants were randomly assigned to one of the three conditions, where they received in-situ support to assist them. They were then asked to evaluate another set of 10 emails, which again included 4 phishing emails, using the provided support.

4.1 Three conditions - Report Type

The three report conditions varied in terms of the type of information displayed and the level of accuracy provided to the participants.

(1) **Control** condition: The standard guidance on avoiding phishing scams (as depicted in Figure 3) was taken from *<Blinded>* University in the UK, with some modifications. The participants were shown the same guidance alongside all emails. This report design matches common guidance given to users by organizations [59].
(2) **Perfect** condition: This condition assesses the impact of an ideal, error-free system. It provides an automated analysis report (Section 3 and Figure 2b) specific to each email. The report presents information extracted from the email (such as from address, actions, sender domain, and destination of links) using automated methods like AI. It also provides the most likely underlying scam based on that information (along with a likelihood score), or if the email is likely safe, it states that. We call this the `Perfect` condition, where the extracted information, analysis, and accuracy percentage displayed are all correct and reliable.

[2]https://www.prolific.com/

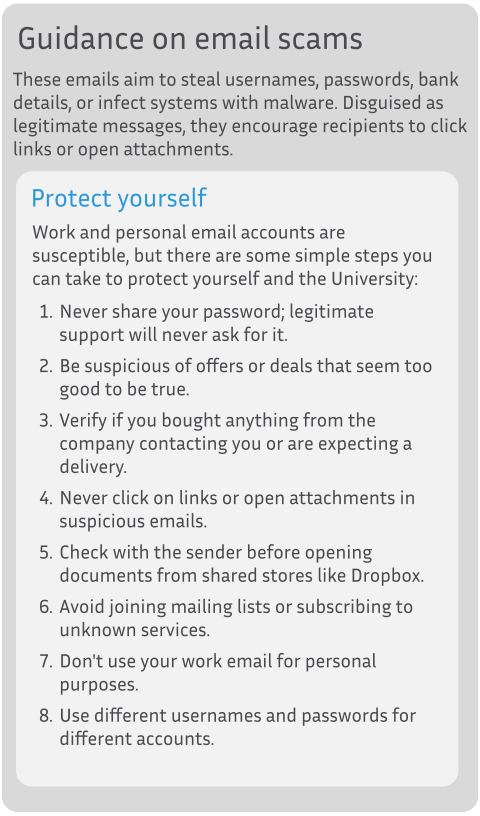


Figure 3: `Control` report - Guidance provided by *<Blinded>* university.

(3) **Realistic** condition: The report in this condition is similar in structure to the `Perfect` condition, but the information provided consisted of errors and the likelihood scores were lower. It is important to note that the participants in this condition were not misled; all facts presented in the reports were real and extracted from the emails. The likely classification section explicitly conveyed uncertainty with phrases like "most likely" and below-100% confidence scores, ensuring participants understood these as predictions, not definitive statements.

4.2 Why is the `Realistic` condition realistic?

The realistic condition is defined as such because the errors we incorporate—*misclassification, parsing errors, and low confidence scores*—reflect common challenges in real-world AI applications, particularly phishing detection [7]. Misclassification occurs due to the probabilistic nature of AI models, where predictions are based on the likelihood of an instance belonging to a specific class. Attackers exploit this by crafting phishing emails that closely resemble legitimate communications, using similar language, branding, and structure. As a result, misclassification is a frequent issue [25, 35, 46, 70]. Despite advancements in detection algorithms [1, 46], no system can achieve 100% accuracy, inevitably leading to false positives and negatives. Additionally, low confidence scores are an inherent

outcome of probabilistic models, particularly when dealing with ambiguous or previously unseen inputs.

The obfuscation techniques used by attackers can also lead to parsing errors [69, 70], which is why we incorporate this type of error. For instance, a common challenge is extracting the correct sender organization name, as attackers often include multiple organization names in their emails, frequently resulting in incorrect identification [71]. Similarly, techniques such as white-on-white text, as shown in Figure 1, introduce noise that complicates parsing and analysis [31]. Hence, the AI errors we simulate in our reports reflect common challenges in phishing detection; however, their exact frequency can only be determined through system deployment and observation.

4.3 Participant Recruitment

We utilized Prolific's screening feature to limit study visibility to participants proficient in English, had completed more than 50 prior submissions, had an approval rate of equal or more than 90%, and were based in the United Kingdom (UK). Phishing email judgements are often highly contextual and require region-specific knowledge, such as recognizing HMRC as the tax authority in the UK or knowing that UK website URLs typically end with a `'.uk'`.

The advertisement was for an "automated email guidance tool" and stated that users would be asked to read 20 emails and perform in-survey actions. The estimated study length was 25 minutes with a compensation of £3.75. Participants were excluded if they did not complete the survey, or revoked their consent. If participants had multiple submissions we kept the first complete response. We also excluded responses not recorded due to technical errors, for which participants received partial compensation and those who answered factual questions about the instructions incorrectly. After eliminating 25 incomplete and invalid responses, our final dataset consisted of 489 participants. The survey took an average of 19.84 minutes to complete, with a median of 17.26 minutes. A post-hoc power analysis confirmed that the sample sizes were sufficient to ensure a robust evaluation of differences between the groups, reducing the likelihood of Type II errors.

4.4 Study Protocol

The study is a between-subjects design with measurements done before guidance was given (round 1) and while guidance was visible (round 2). The study was administered using Qualtrics[3] and the participants were recruited from Prolific. The study was reviewed and approved through our institution's Research Ethics Process. A full version of the survey is available in supplementary material.

Instructions. Participants were first asked to read and agree to a consent form and were then provided instructions explaining that they would need to differentiate between scam and legitimate emails. The emails were modified to be sent to "Emily Morrison," a student at a *<Blinded>* university and specific details such as the university name and Emily's email address were provided. Participants were then asked what Emily's university name and email address were to ensure they understood the information. The "Emily" character is used because emails are highly contextual. While none of the emails presented relied on the user knowing details about Emily, some of the legitimate emails might appear suspicious if the intended recipient is unknown.

Participants were informed that they would be presented with 10 emails, some of which would be scams. They were asked to assess the likelihood that each email is a scam or legitimate based on the content and primary links provided in the screenshots. It was made clear to participants that it was not necessary to use external sources such as Google to complete the tasks.

Round 1. All participants in Round 1 were shown the same 10 emails in randomized order. For each email, participants were shown a screenshot of the email and asked two questions: if it is a scam (no, yes) and their confidence in their answer ('Not at all confident' to 'Extremely confident', 5-point Likert). A benign email in the set was designed as an attention check and asked participants to select "Moderately confident", hence confidence scores are based on 9 emails in round 1.

Report Explanation. Participants were randomly assigned to one of the three conditions and given slightly different instructions depending on their condition. They were informed that they would again be presented with 10 emails, some of which were malicious. They were shown a short use case comic showing a person looking at a suspicious email and clicking a button to get guidance. The `Control` condition was told that the guidance was from the University. The `Perfect` and `Realistic` conditions stated that the report contained an analysis generated by an AI model specific to the email.

Round 2. Similar to Round 1, participants were shown 10 emails in randomized order, 4 of which were malicious. For each email, they were shown a screenshot of the email and the associated report. `Control` participants always saw the same set of guidance (Figure 3), while the other two conditions had reports that were unique to each email (similar to Figure 2b). Participants were asked three questions: if it was a scam, their confidence in their answer, and which part of the report/guidance helped them most when answering. Note that due to an error, one of the phishing emails was not correctly displayed to a percentage of the participants. Out of an abundance of caution, we removed the email judgement from the dataset, so Round 2 data is computed out of 9 emails even though 10 were shown.

Concluding Questions. The System Usability Scale (SUS) [13] is a widely recognized general tool for assessing the usability of a product or service consisting of a ten-item questionnaire. First, we asked SUS questions by replacing the word 'system' with the word 'report' in the scale to avoid participant confusion. We also asked if they thought such a tool would improve email security, confidence with the report compared to without it, self-report ability to identify report inaccuracies, if they would recommend such a report to a friend, and invited optional feedback on how to improve the report in a free text box. For demographics, we asked participants their age, gender, and IT or security work experience. Finally, we asked about their security practices using the proactive awareness sub-scale of the Security Behavior Intentions Scale (SeBIS) [30]. This sub-scale assesses the participants' proactive awareness of their online browsing habits. The questions aim to measure how

[3]https://www.qualtrics.com/

frequently individuals engage in security-conscious behaviours, such as verifying the destination of links before clicking, recognizing secure website indicators, and taking action upon discovering security issues.

4.5 Emails Presented

During the study, participants viewed 20 emails in two sets of 10. In each set, 6 emails were benign (safe) and 4 were phishing emails. All emails were based on real and phishing emails sent to *<Blinded>* University inboxes of Informatics students and staff. Some phishing emails were also based on emails found in the Nazario Phishing Dataset [61], a publicly available collection of hand-screened phishing messages by Jose Nazario.

We selected phishing emails based on the 'Factors of Phishing Sophistication' scale proposed by Kersten *et al.* [45]. This scale defines four factors of phishing sophistication that directly influence the believability of a phishing email: *Technical (T), Contextual (C), Language and Tone (Lg), and Layout (Ly).* Each phishing email in each round was designed to be sophisticated in one factor. The Technical (T) factor focuses on technical elements in the email, such as the sender's name, domain, and links. Contextual (C) factor looks at how well the phishing email's pretext aligns with the target's specific situation. The Language and Tone (Lg) factor examines the appropriateness of the language and tone used in the email in relation to the target, while Layout (Ly) evaluates how closely the email's visual presentation resembles what the target would expect from a legitimate message. This scale was used to ensure both rounds (without and with guidance) presented participants with similar phishing emails. This approach ensured a fair comparison, avoiding any imbalance between easier and harder-to-detect phishing emails across the rounds.

Once we selected the emails, we manually edited the files to remove any personally identifying information and changed references to people, but kept the original links from the emails. We then captured images of the edited emails for the survey. In all email images, the target of the main hyperlink (the one intended to be clicked on) was shown at the bottom of the image in the status bar and indicated by the hovering pointer, as shown in Figure 2a.

5 RESULTS AND ANALYSIS

5.1 Participant Demographics

Demographics were drawn from Prolific-provided data and the Qualtrics survey. Table 1 shows the demographics distribution of our participants. 254 participants (51.94%) identified as female, 219 (44.79%) as male, 2 (0.41%) preferred not to disclose their gender, and 14 participants (2.86%) did not allow demographic data to be downloaded from Prolific. A chi-square test of independence showed no significant association between gender and report conditions, $\chi^2(6, N = 489) = 6.999, p = 0.32$. In terms of age, the oldest participant was 80 years old, and the youngest was 18, with an average participant age of approximately 40. Age data for 14 participants was unavailable from Prolific. An ANOVA test shows no statistically significant difference in age distribution across the different conditions ($F = 0.63$, $p = 0.53$). When asked about their work experience and education related to Information Technology (IT) or computer security, 49.90% of the participants indicated no experience, 30.06% reported 'A little', and 20.04% indicated that they had experience. There was no significant association between technology experience and the conditions, $\chi^2(4, N = 489) = 8.35, p = 0.08$. The SeBIS proactive awareness subscale [30] had a mean of 20.30 out of a possible 25, and a standard deviation of 3.03, with scores ranging from a minimum of 9 to a maximum of 25. An ANOVA test showed no significant difference in proactive awareness scores across the conditions, $F(2, N = 489) = 2.12, p = 0.12$. We assessed data normality using the Shapiro-Wilk test [75] and evaluated the homogeneity of variance through box plots and Levene's test [37].

Table 1: Demographic information: participant age, gender, experience or education in information technology or computer security.

Demographics	**Distribution**
Gender	
Male	219 (44.78%)
Female	254 (51.94%)
Not Available	16 (3.27%)
Age	
< 18	0 (0%)
$18 - 24$	52 (10.63%)
$25 - 34$	141 (28.83%)
$35 - 44$	129 (26.38%)
$45 - 54$	70 (14.31%)
55+	83 (16.97%)
Not Available	14 (2.86%)
Technology Experience	
Yes	98 (20.04%)
A Little	147 (30.06%)
No	244 (49.90%)

5.2 Impact of Guidance

RQ1 examines how real-time guidance affects users' accuracy and confidence in identifying phishing emails. Participants answered questions in round 1 (R1) without guidance and in round 2 (R2) with guidance. The independent variable is the stage (R1 vs. R2), and the dependent variables are accuracy and confidence. Accuracy was measured by asking whether each email was a scam (Yes/No), and confidence was rated on a five-point confidence Likert scale. Since the data was non-parametric, as shown by the Shapiro-Wilk test, a Wilcoxon Signed-Rank Test [23] was used to analyze changes in accuracy and confidence between R1 and R2. Table 2 summarizes the scores by condition and round.

Analysis of Accuracy Scores: The accuracy score for each participant per round was calculated as the mean of correct judgments, producing a value between 0 and 1. The mean accuracy score for all participants across the three conditions increased from 0.78 without guidance to 0.84 with guidance (Table 2), which was a statistically significant change ($W = 35183.5$, $p < 0.001$). Furthermore, we used a one-sided Wilcoxon test to check for *improvements* in each

condition. The Perfect condition showed the most statistically significant improvement, with mean accuracy increasing from 0.78 to 0.93 ($p < 0.001$). The Control condition showed a smaller improvement from 0.79 to 0.83, which was also significant ($p = 0.0131$) In contrast, the test for the Realistic condition shows no statistically significant improvement ($p = 0.6057$) as the accuracy was similar in both stages (0.78).

Analysis of Confidence Scores: The confidence score for each participant was computed by taking a mean of their responses to the five-point confidence Likert scale, where 5 indicated "Extremely confident" and 1 indicated "Not at all confident." The mean confidence score for all participants across the three conditions increased from 3.60 without guidance to 3.83 with guidance, which was a statistically significant change ($W = 24205.0$, $p < 0.001$). Again, we used a one-sided Wilcoxon test to test for improvements in confidence scores in each condition. The Perfect condition saw a statistically significant improvement ($p < 0.001$) of mean confidence from 3.60 to 4.11. For the Realistic group although the accuracy did not change, there was a statistically significant improvement in confidence from 3.52 to 3.74 ($p < 0.001$). In contrast, the Control condition showed no significant improvement in confidence scores ($p = 0.8401$), and in fact, there was a slight decrease in mean confidence (3.68 to 3.64).

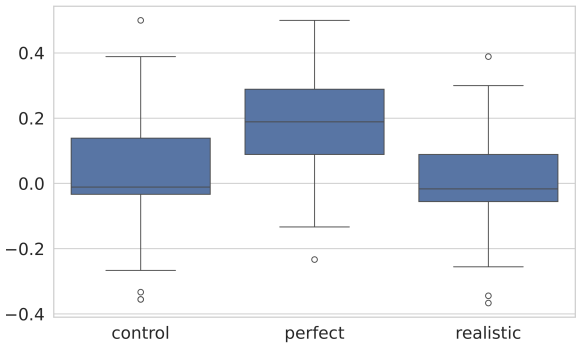
The results indicate that having a tailored report with relevant advice based on specific emails significantly improved participants' performance. A closer analysis of the data revealed that 265 out of 489 participants improved their accuracy in the second round, while 313 participants demonstrated increased confidence. Such improvements in a real-world organization could greatly reduce security risks by enhancing employees' ability to detect phishing attempts. Furthermore, the increase in confidence could lead to higher reporting rates, fostering a more proactive security culture.

Table 2: Descriptive statistics (mean (M) and standard deviation (SD)) for accuracy and confidence scores by condition (Overall, Control, Perfect, Realistic) and stages; without guidance (R1) and with guidance (R2). Accuracy scores are on a $0-1$ scale and confidence score are on $1-5$ scale.

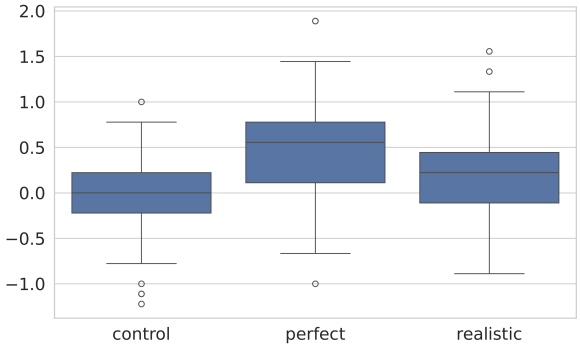
Condition	Stage	M_{acc}	SD_{acc}	M_{conf}	SD_{conf}
Overall	R1	0.78	0.13	3.60	0.59
	R2	0.84	0.14	3.83	0.62
Control	R1	0.79	0.14	3.68	0.59
	R2	0.83	0.15	3.64	0.63
Perfect	R1	0.78	0.13	3.60	0.60
	R2	0.93	0.10	4.11	0.58
Realistic	R1	0.78	0.12	3.52	0.57
	R2	0.78	0.12	3.74	0.55

5.3 R2 & RQ3: Impact of Types of Guidance

To study the impact of the three report conditions (Control, Perfect, and Realistic), we compare the change in accuracy (Δacc) and confidence ($\Delta conf$) between the three conditions. We first calculate



(a) Changes in accuracy scores from round 1 to round 2.



(b) Changes in confidence scores from round 1 to round 2.

Figure 4: Changes in accuracy/confidence scores in different rounds

the difference between the average Round 2 and Round 1 score per participant for both accuracy and confidence and then use a one-way analysis of variance test (or one-way ANOVA) [54], which resulted in a statistically significant difference for both accuracy ($p < 0.0001$) and confidence ($p < 0.0001$). We performed a post-hoc analysis between all three conditions using Tukey's honestly significant difference test (Tukey's HSD). All three pairs showed significant differences in confidence change ($p < 0.001$). Accuracy change was statistically significant between Control-Perfect conditions ($p < 0.001$) and Perfect-Realistic conditions ($p < 0.001$). However, there was no significant difference between the Control-Realistic conditions ($p = 0.0795$), indicating similarities in the impact of generic guidance and imperfect guidance.

In terms of accuracy change, the Perfect condition showed the largest improvement ($M_{\Delta acc} = 0.153$), while Control showed a small improvement ($M_{\Delta acc} = 0.036$), and Realistic resulted in virtually no change ($M_{\Delta acc} = 0.001$). For change in confidence, the Perfect condition shows the most substantial increase ($M_{\Delta conf} = 0.504$) compared to the smaller increases in Realistic ($M_{\Delta conf} = 0.220$) and a slight decrease in Control ($M_{\Delta conf} = -0.033$). This suggests that participants felt significantly more confident when provided with perfect guidance. The Perfect condition outperforms both Control and Realistic in both accuracy and confidence improvements, with statistically significant differences observed between Control-Perfect and Perfect-Realistic conditions, but no significant difference in accuracy between Control

and Realistic conditions. Overall, the findings suggest that the Perfect condition is most effective in enhancing both accuracy and confidence.

Table 3: Mean (M) and Standard Deviation (SD) of Accuracy and Confidence changes across the three conditions. Mean accuracy change ranges from $0-1$ and mean confidence change ranges from $1-5$.

Condition	$M_{\Delta acc}$	$SD_{\Delta acc}$	$M_{\Delta conf}$	$SD_{\Delta conf}$
Control	0.04	0.15	-0.03	0.38
Perfect	0.15	0.15	0.50	0.48
Realistic	0.00	0.13	0.22	0.41

An in-depth analysis of accuracy in Round 2 (with guidance) showed that Perfect condition had a mean accuracy of 93.1% and a median of 100%, 93/160 participants in this condition correctly answered all emails. In contrast, Realistic had a mean of 78.3% with only 10/166 answering all emails correctly and Control had a mean of 82.9% with 35/163 answering all emails correctly. In R1, participants had more trouble with benign emails (74.2% correctly judged) than phishing emails (84.8% correct). In R2 Control (benign=78.4%, phish=88.5%) continued this trend, but Realistic (benign=80.1%, phish=75.9%) does better on benign than phishing and Perfect (benign=93.0%, phish=93.4%) saw nearly identical scores for phishing and benign. This suggests that having a contextual report, even an inaccurate one, helped people better make decisions about safe emails.

Confidence follows a somewhat similar pattern. In R1, in only 54.1% of email judgements did the participant indicate they were "Extremely" or "Very" confident indicating *high* confidence. Similar to accuracy, they had higher confidence on phishing emails (59.8%) than benign emails (50.3%). In R2, the Perfect condition marked 75.4% of judgements as high confidence. By comparison, Control saw limited improvement with only 56.2% marked as high confidence and Realistic saw only limited improvement with 61.2% marked as high confidence. Seeing tailored reports did increase confidence compared to generic guidance.

Participants' confidence partially aligned with their accuracy. In R1, participants who accurately judged the email showed high confidence in 57.6% of the cases. Conversely, they indicated high confidence in 41.3% of the incorrectly judged cases. This suggests that they performed better than guessing, but their confidence was not a strong indicator of accuracy. In R2, again the Perfect condition did the best with 77.8% of the correct judgements having high confidence and 42.9% of incorrect judgements indicating high accuracy. Realistic saw marginal improvement over R1 (66.2% of correct, 43.1% of incorrect) and Control marked their incorrect judgements with the lowest confidence (60.9% of correct, 33.5% of incorrect). Overall, confidence in judgments was not particularly high, even when participants were provided with accurate reports and suggestions. This indicates that participants are aware that such reports may not always be accurate, leading them to not always be confident in their responses.

5.4 Impact of different parts of the report on participants' decisions

In R2 after participants judged each email, they were asked what part of the guidance (Control) or report (Perfect, Realistic) most helped when making their decision. In the Control group, the most common response was "I did not use the report to decide" (454 out of 1467 responses), with participants particularly disregarding the report for benign emails (3 out 5 emails). For phishing emails, participants frequently selected "Never click on links or open attachments in suspicious emails." In the Perfect group, "Sender Domain" was the most common response (516 out of 1440 responses) and for six out of nine emails. This was also the most commonly selected choice for benign emails (3 out of 6 emails). For phishing emails, participants commonly selected "Link destination". The Realistic group had the most frequent response, "All of the above" (518 out of 1494 responses), although "Sender Domain" was the most common choice for five emails, including phishing emails. For benign emails, "All of the above" was the most selected. The report format provided goes beyond simply displaying the sender domain, it also emphasizes whether the domain matches the extracted organization name ("Email claims to be from"). Notably, the source of the email plays a significant role in influencing users' decisions. For both Realistic and Perfect groups, this was a common choice, especially for phishing emails. This is probably because the discrepancies between the sender's domain and the claimed organization are strong indicators of fraud.

In the Realistic group, there was a tendency to use a holistic approach, with "All of the above" being the most common choice for half of the emails, especially in the benign emails. In contrast, the Control group demonstrates a high level of disengagement, with participants frequently selecting "I did not use the report to decide," particularly for benign emails (4 out of 6). This suggests that participants in the Control group have relied more on their own judgment rather than the report, which was the same generic advice given for every email. The Perfect group consistently relied on "Sender Domain" across most emails, highlighting the importance of this specific cue in their decision-making.

5.5 Impact of Inaccuracies: Two emails under scrutiny

For the Realistic condition, we created reports that simulated common errors made by AI systems (4.2). This included instances where phishing emails were incorrectly classified as safe, and benign emails as phishing. This reflects the probabilistic nature of AI. We provided supporting evidence, such as incorrectly parsed link destinations and low likelihood percentages. This allowed us to study the impact of imperfect guidance on user performance. Table 4 shows the percentage of participants who identified the email correctly as phishing (P) or benign (B) in Round 2. Emails whose reports contained inaccurate guidance in Realistic condition are marked by bold and an 'I', along with the induced error. We now discuss in detail two emails that showed very interesting response patterns across the groups. The response patterns for all of the inaccurate reports are detailed in Appendix 7.

Table 4: Percentage of participants who identified the email correctly as phishing (P) or benign (B) in Round 2. The inaccurate guidance in the Realistic condition is marked by bold and an (I). Email presentation order was randomized, numbers are for easy reference.

	P/B	Control (%)	Perfect (%)	Realistic (%)
R2_1	P	66.87	81.88	**33.13** (I: Action, Recommendation)
R2_2	P	91.41	95.63	**96.99** (I: Link destination)
R2_3	P	98.16	98.75	96.39
R2_4	P	97.55	97.50	**77.11** (I: Recommendation)
R2_6	B	87.73	97.50	96.99
R2_7	B	80.37	96.25	93.98
R2_8	B	85.28	96.25	**90.96** (I: Recommendation correct but % low)
R2_9	B	88.34	96.88	96.39
R2_10	B	50.31	78.13	**22.2** (I: Recommendation)

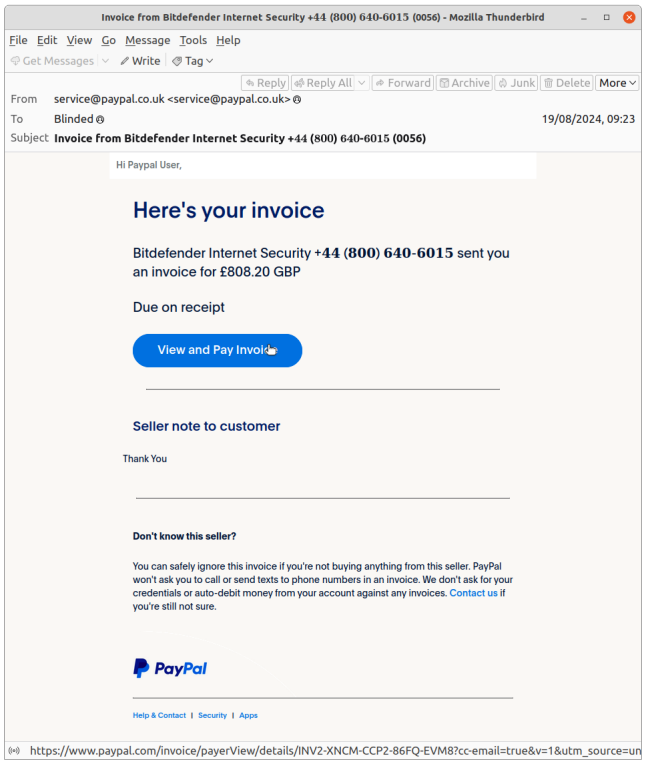


Figure 5: The PayPal email that was shown to participants in round 2 (R2_1). The sender domain is visible in the header and the URL is shown at the bottom.

R2 Email 1: High-Technology Phishing Email. The first email (Figure 5) was designed to be technically sophisticated but low in contextual, linguistic, and layout cues, following the phishing sophistication scale. It appeared to be from PayPal, with a valid sender domain ('paypal.co.uk') and a button linked to a non-existent page at 'paypal.com,' a valid PayPal domain. The trick was the link leading nowhere, prompting the recipient to call a phone number. This email is a good example of where an AI might realistically make an error. The `Realistic` report erroneously detected the intended action to be 'click' and consequently marked the email safe because a user clicking on the links would indeed be safe. However, since the linked page does not exist, a user is likely to then call the phone number which is not safe. For balance, a similar email claiming to be from WeTransfer was also shown in R1 with only 68.3% of participants correctly detecting it.

The subtlety of the phone number vs link destination is likely what made this email the second most challenging in round 2 for all conditions. This is an example of a phishing attack which closely mimics or sometimes is even legitimately sent by a real service and the phishing attack is in the message itself rather than the sender or links. The `Realistic` condition was clearly impacted by the incorrect report with only 33.1% correctly judging the email. Even the `Perfect` condition who had good guidance only correctly answered 81.9% of the time. Those who were correct indicated that the classification of scam impacted their decision, while those who were wrong often indicated the destination of the link. The `Control` condition with generic guidance only answered 66.9% correctly and when asked what most impacted their answer they indicated the generic guidance to "verify if you bought anything from the company contacting you or are expecting a delivery".

R2 Email 10: Benign Email about an Internship Opportunity. This was a benign email from Standard Life Investments about a summer internship (Figure 6). The sender domain was within the university and the link was a shortened URL (bit.ly). Although the email was sent from an internal university account, this email had the lowest performance in all three report conditions. This email is a good example of a well-meaning benign email that happens to have many features associated with phishing such as: being from an unfamiliar sender, using a bit.ly link, talking about an opportunity, and including lots of €-symbols in the bullet points. It is important to remember that not all email writers are trained in writing genuine-looking emails. Groups, such as students, that are seeking job opportunities have an earnest need to evaluate such offers rather than delete them "just in case". Consequently, this type of email is particularly challenging.

50.31% of the `Control` group correctly identified the email as safe, most of whom said they did not use the generic guidance to make this choice. On the other hand, most of those who identified it as phishing cited the guidance to 'Never click on links or open

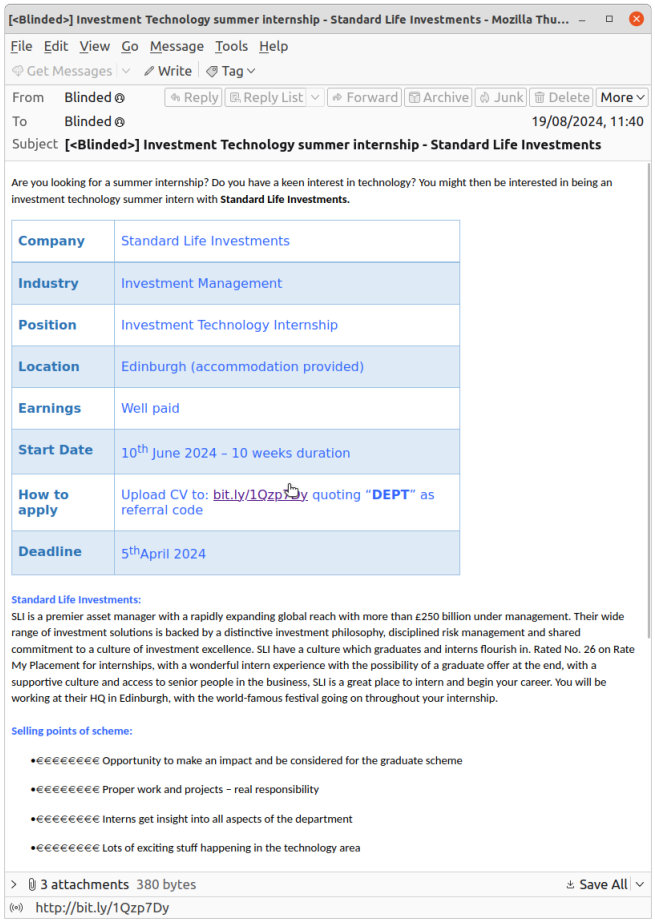


Figure 6: The internship email that was shown to participants in round 2 (R2_10). The sender domain is visible in the header and the URL is shown at the bottom.

attachments in suspicious emails' as the reason. The `Perfect` group were told this email was 95% likely to be safe. 78.13% correctly identified it as safe with the sender domain as the most commonly cited evidence, while 21.87% classified it as phishing and mostly cited link destination which the report identified as "NA" and unverified as the reason. The `Realistic` group's report classified the email as a financial scam with a low percentage (65%). This was the only difference in the reports provided to the two groups, and the evidence shown was identical. However, only 22.2% of participants correctly identified the email as safe. The most common reason cited was the sender domain or that the participant did not use the report (suggesting their own judgement). A large number of participants (77.8%) followed the provided guidance and used the link destination to identify the email as a phishing email.

Relying only on sender and URL information is not enough. These two emails show that systems that rely heavily on sender information and URLs are problematic. Legitimate domains often spoofed or redirected through, could cause URL and sender-based warning systems to fail, and users would then struggle to identify the scam. Furthermore, shortened URLs are hard to judge as they are commonly used legitimately but are also a common phishing tactic. These cases highlight the need to go beyond just sender and URL information and incorporate additional cues in phishing guidance systems.

Will users blindly follow? In the `Perfect` condition, 89 participants followed the system's recommendations for all emails in round 2. This indicates that users do not blindly rely on the system even when it is accurate. In contrast, in the `Realistic` condition, where the reports included errors, only 16 participants followed the system's recommendations for all emails. This indicates that in the presence of errors, many users chose to deviate from the system's classifications, demonstrating an awareness of potential inaccuracies and a willingness to rely on their own judgment.

Supporting evidence is important. Comparing response patterns for two phishing emails in round 2, R2_1 and R2_4, reveals key insights (Table 4). Both were misclassified as safe by the system, but participants responded differently. In R2_1, where both the domain and sender appeared safe, many participants trusted the system's recommendation. Conversely, in R2_4, despite the URL pointing to a legitimate government domain, the suspicious sender domain led fewer participants to follow the system's advice. This highlights that user decision depends not only on the classification but also on the supporting evidence provided.

5.6 Post-survey Analysis to Understand the Report Usage

At the end of the survey, we asked several questions aimed at understanding the participants' views on the report usage.

5.6.1 System Usability Scale [13]. As mentioned in Section 4.4, we adapted the SUS by changing the word "system" to "report" to clarify what the participant was responding to regards to for assessing the usability of the guidances provided. The mean SUS scores for the three conditions were `Control`: 52.39, `Perfect`: 55.38, and `Realistic`: 54.49. Both `Perfect` and `Realistic` had higher scores than `Control` suggesting that a structured report design is an improvement over static guidance. That said, all the SUS scores fall well below the benchmark of 70 recommended to judge the interface as usable, and we will work on improving the design of the guidances provided as part of our future work.

5.6.2 Self-reported views. When asked if the participants thought that the tool would improve their email security, 18.81% responded with 'Definitely yes' and 46.21% responded with 'Probably yes'. Additionally, when asked if they would recommend the tool to a friend or colleague, 45.39% responded with 'Probably yes' and 15.33% responded with 'Definitely yes', reflecting their belief that the tool can be a valuable asset to others as well. When surveyed, 79.4% of participants in the `Perfect` condition indicated that they believed the tool would enhance their security, responding with either "Probably" or "Definitely yes." Furthermore, 75.6% said they would likely recommend the tool to a friend or colleague. These results highlight the crucial role of immediate, high-quality guidance in improving email security.

6 DISCUSSION

Attackers are constantly developing new sophisticated phishing emails, making them more difficult for non-expert users to detect. Supporting users in identifying these malicious emails and links has been a crucial area of research for many years. Most users have access to some form of advice or guidance, typically offered by their organizations through phishing training programs [38, 49, 76, 92] or phishing reporting systems [5, 51]. However, research has shown that the effectiveness of phishing training diminishes over time [9, 19], while the process of reporting and waiting for assistance can be time-consuming, potentially causing users to act impulsively out of curiosity [88] or urgency [17]. This leaves the users vulnerable to phishing attacks in the critical moments when they need immediate help, highlighting the importance of real-time guidance. But this raises the question of what this guidance should contain. While most organizations have standard generic guidance, an ideal approach would be to provide tailored guidance specific to the recipient [29, 36, 41]. An efficient way is to generate guidance automatically to help the users in a timely manner.

There are many ways to generate phishing guidance. Due to the growing advancements in AI (such as the large language models), one natural choice would be to leverage AI to generate context-specific guidance. Such AI-assisted decision-making is growing in research and practice (e.g., [48, 80, 89]). However, a significant downside of implementing AI-based systems for phishing guidance is the potential impact of errors or inaccuracies in AI-generated reports, which are inevitable given the nature of AI algorithms [7, 16]. Despite this, an AI-assisted approach still holds promise for providing users with (in)accurate guidance, thus supporting them to effectively detect and respond to potential phishing threats. In our study, we examine the impact of such a hypothetical guidance system on user phishing detection ability.

Our findings show that access to real-time guidance can significantly improve both accuracy and confidence in phishing detection. Even in the generic, non-tailored condition, user performance improved, with 74 participants demonstrating better accuracy. This suggests that even basic, non-personalized guidance delivered at the moment of risk can have a positive impact on decision-making. This could contribute to creating safer organizations by reducing the number of phishing victims. The boost in confidence also leads to higher reporting rates, which are crucial for proactive threat mitigation.

6.1 Make it Relevant: The Importance of Customized Email Guidance

The findings from our study highlight the critical difference between customized and generic email guidance on users' phishing detection capabilities and confidence. While the `Perfect` group received highly tailored guidance for each email, the `Control` group received only generic guidance, and the participants did not fare as well. The `Control` group experienced a much smaller improvement in their accuracy in making decisions. The generic guidance, similar to that provided by many organizations, was informative but did not address the specific characteristics of each email. As a result, users in the `Control` group found it more challenging to apply the provided guidance to specific instances, leading to confusion and reduced effectiveness. When asked what part of the guidance they used, many indicated: 'I did not use the report to decide' compared to the other two conditions. This result is further supported by the feedback received from the participants in this group where many complained that the guidance was too generic; *e.g., "report seemed generic and didn't highlight clues from the email"; or "more specific details relevant to the email shown instead of showing the generic points".* In Section 5.5, we show how commonly used phishing tactics such as domain spoofing, redirection, and the use of shortened URLs [4] can significantly influence users' decisions. A system that relies solely on domains and URLs (the `Realistic` group) would fail in such cases, leading many users to fall for the phishing attempt. Generic guidance could go either way, but a perfect guidance system based not only on domains but also on incorporating other cues in emails can significantly improve performance.

The impact of this generic guidance was evident in the participants' perceptions of the tool's value. When asked if they thought the tool would improve their security, only 46.62% of participants in the `Control` group responded with "Probably" or "Definitely yes". This response rate is significantly lower than that of the `Perfect` group, suggesting that generic guidance does not inspire the same level of confidence in users regarding their ability to protect themselves against phishing attacks. Similarly, only 44.78% of the `Control` group participants indicated that they would recommend the guidance to a friend or colleague. This lower recommendation rate points to a perceived lack of effectiveness and relevance in the guidance provided.

6.2 Judging Phishing Under Uncertainty: Impact of Inaccuracies

Automated systems are not perfect, inaccuracies and errors are inevitable. Understanding the impact of these errors is crucial. In our study, we examined how wrongly extracted information or incorrect email classifications affected user performance. We did this through the `Realistic` condition which received guidance similar to the `Perfect` condition but with intentional inaccuracies introduced. While this group saw no improvement in detection accuracy, they still experienced an improvement in confidence, much less compared to the `Perfect` group. When participants were asked if they believed the automated guidance would improve their security, only 69.28% responded with "Probably" or "Definitely yes". Similarly, when participants were asked if they would recommend the automated guidance to a friend or colleague, only 62.05% responded with "Probably" or "Definitely yes". While still a majority, these numbers are lower than the `Perfect` group indicating that inaccuracies not only affect the user's trust in the automated guidance but also their willingness to advocate for its use by others.

In Section 5.5, we discuss in detail emails with "realistic" reports that contain errors. The analysis shows that the accuracy of automated email guidance systems can significantly impact user decision-making. When such systems provide incorrect classifications, especially for common phishing tactics like spoofed domains or shortened URLs, users are more likely to misjudge the safety of an email, resulting in serious consequences. However, if the overall high-level classification is correct, even when some evidence is wrongly extracted, user performance remains largely unaffected,

as they rely on the final recommendation. Conversely, if users recognize that the automated guidance is presenting false or incorrectly extracted evidence, it can still lead to a slight reduction in confidence. Additionally, low classification percentages, which are typical in probabilistic AI models, tend to diminish user trust and confidence in the system's recommendations.

The feedback from the `Realistic` group participants showed several concerns regarding the accuracy. Many participants felt that the inconsistencies identified in the automated guidance decreased their trust in its judgments. For instance, some users mentioned that the automated guidance sometimes showed emails as safe even when certain elements appeared suspicious, which created confusion and reduced confidence in the guidance's reliability, *e.g.,"I didn't feel I could trust the report because some of the judgements it made were not consistent"; or "...some percentages were say 55% which didn't fill me with much confidence"; or "Fewer inconsistencies in the report..."*. We acknowledge that the "inaccurate" guidance in the realistic condition may have caused some frustration, confusion, or reduced confidence in participants' decision-making. Although this approach was essential for studying the impact, it may have led to confusion regarding email cues and the participants' own judgments. To mitigate this, we explicitly conveyed uncertainty using phrases like "most likely" and confidence scores below 100%, to help participants understand that these are predictions rather than definitive statements. They were encouraged to use their judgment alongside the reports to make their final decisions. Additionally, an analysis of their written feedback indicates that they recognized the system's imperfections. While some participants may have been negatively affected, we believe that most participants understand that computers can make mistakes, as even seemingly basic tools like spell check often need human judgment to catch inaccuracies.

6.3 No Guidance vs Generic Guidance

One of the most notable observations from our study is that the `Control` condition with generic guidance can outperform tailored but inaccurate guidance. Interestingly, the `Control` condition showed a statistically significant improvement in accuracy, while the `Realistic` condition exhibited no accuracy gains. As demonstrated in Table 4, performance on two specific emails (email 1 and email 10) where the `Realistic` condition saw errors, resulted in substantially lower accuracy compared to the other conditions. In other words, inaccurate advice can have a disproportionately negative impact on decision-making. Despite these two poorly performing emails (*R*2_1 and *R*2_10), the performance of the `Realistic` condition is better than the `Control` and closer to the `Perfect` condition in most of the cases. This observation highlights that in instances where the guidance system makes high-level misclassifications, even generic guidance proves more beneficial by avoiding errors that lead users astray.

Our findings highlight the need to rigorously evaluate and address potential inaccuracies in automated guidance systems before deployment. The results raise the question of whether these systems might ultimately cause more harm than benefit compared to traditional, generic systems. To counter this, we propose implementing a threshold likelihood score, below which the system would remain inconclusive and provide generic guidance rather than definitive classifications, reducing the risk of misclassification. Additionally, we advocate for the development of a comprehensive scam likelihood score that integrates multiple dimensions of email analysis, moving beyond reliance on isolated factors such as URLs. This requires an in-depth investigation into the cues IT staff typically utilize when addressing phishing threats, ensuring the score aligns with real-world practices and decision-making. Finally, we suggest adopting a "safer-than-sorry" approach, where even minimal indicators of phishing trigger scam warnings, prioritizing user safety and minimizing potential harm. Together, these strategies aim to enhance the reliability and effectiveness of automated phishing guidance systems.

6.4 Implications of the Response Patterns

Insights from the `Realistic` condition highlight the importance of evidence presented in guidance reports. For example, R2_1 and R2_4 were both phishing emails misclassified as safe, yet participants responded differently: 77.87% considered R2_1 safe due to its legitimate sender and URL, while only 32.89% considered R2_4 safe, as it had a suspicious sender despite a legitimate URL. *This shows that participants rely not only on classifications but also on the supporting evidence, highlighting the need for guidance systems to include clear explanatory information.* For R2_10, a legitimate email from an internal sender, elements like a short URL and unusual formatting led many participants in the generic and realistic conditions to misclassify it as phishing, resulting in false positives that waste organizational resources. T*o address this, we suggest implementing clear formatting guidelines for organizational emails, especially for mass communications, to avoid triggering red flags.* Interestingly, in the `Perfect` condition, more participants trusted the 'safe' guidance for R2_10 despite the suspicious elements. Comparing R2_4 and R2_10 reveals that when an email contains both safe and suspicious elements (sender or URL), users often experience uncertainty about how to interpret the guidance. In such cases, they are more likely to follow the system's recommendation, reinforcing the importance of providing accurate and well-supported evidence in guidance systems.

6.5 Limitations

We recruited participants for our study using Prolific, which limits the generalizability of the results. Prolific participants are experienced survey takers and have higher levels of technology use relative to the general population, which likely influenced the results. However, recent research has shown that data collected through Prolific is generally representative of questions about user perceptions and experiences compared to other platforms [83]. Additionally, we only allowed UK residents to participate as the contextual nature of our emails required a certain level of knowledge about life in the UK. This decision limits the generalizability of the results to other countries, particularly those that may not communicate the same way over email.

The report template used in our study was created based on our understanding of the phishing problem and insights from prior research on phishing education and awareness. The design was intended to be simple and easy to understand, allowing participants to quickly grasp the information provided; and in this study, our

focus was on the content of the guidance. However, we did not conduct a formal pre-study to evaluate the design before its use in the survey. Consequently, there may have been missed opportunities to improve its effectiveness and user engagement through iterative testing and feedback. We aimed to model scenarios that reflect plausible inaccuracies users might encounter, based on existing research and practical observations. However, we acknowledge that real-world inaccuracies may vary in type and frequency and are hard to replicate without actual deployment.

Another important limitation is the study's ecological validity. We provided email analysis alongside each email, whereas in a real email client, such a feature would require user-initiated action, like clicking a button. Although we used a comic to explain this scenario in the instructions, it does not fully capture natural user interaction. Participants may therefore have been in a different mindset than the intended users. We used the email classification task to measure accuracy, acknowledging its limitations [33, 73, 85]. Participants lacked the full context of the email recipient, such as familiarity with the sender or typical email content, which are crucial in real-world phishing detection. Additionally, the study's focus on email identification as a primary task does not reflect the multitasking environment of real-world users, where limited attention could impact their decisions. Despite these limitations, we believe that the insights from this study will contribute to the development of improved systems and give readers a clear understanding of the advantages and disadvantages of AI-assisted phishing guidance systems.

7 CONCLUSION

Providing relevant and actionable advice to users about phishing emails is a challenging task, especially with automation. Commonly used banner warnings or URL-based warnings have limitations, and the majority of guidance is generic. The emails that end-users encounter in their inboxes have likely already passed through advanced filters and checks and would require contextual knowledge that users have to make a good judgement. In this study, we investigated the impact of real-time guidance, based on features extracted from emails, on user performance (accuracy and confidence) through an online survey of 489 participants. We compared these results to the impact of generic guidance, similar to that provided in many organizations. Additionally, we examined the impact of inaccurate information presented in the reports on user performance. By measuring user accuracy and confidence in phishing detection with and without guidance, we found that real-time guidance significantly improves both accuracy and confidence, particularly in the `Perfect` condition where highly accurate tailored advice was provided. However, inaccuracies in the guidance, especially incorrect scam classifications, led many users to make wrong decisions, reducing overall accuracy. In terms of confidence, tailored reports (`Perfect` and `Realistic`) increased users' confidence, while generic guidance minimally impacted it, likely due to a mismatch between the advice provided and the context of the email. This work not only evaluates a novel type of user assistance system that provides real-time contextual guidance but also pioneers the study of how errors in automated systems impact user performance. By examining how inaccuracies influence user behaviour, we contribute new insights into AI-assisted decision-making in phishing. In future work, we plan to investigate the long-term effects of such phishing guidance, particularly whether repeated use of automated guidance leads to user education, and if using such systems regularly would improve user knowledge and skills.

ACKNOWLEDGMENTS

This work has been partially supported by Google through a Google Research Award and by the Natural Sciences and Engineering Research Council of Canada (NSERC) under award number RGPIN-2024-06737. The authors would like to express their gratitude for this support, which has been instrumental in this research.

REFERENCES

[1] Saeed Abu-Nimeh, Dario Nappa, Xinlei Wang, and Suku Nair. 2007. A Comparison of Machine Learning Techniques for Phishing Detection. In *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*. Association for Computing Machinery, New York, NY, USA, 60–69. https://doi.org/10.1145/1299015.1299021

[2] Ujué Agudo, Karlos G Liberal, Miren Arrese, and Helena Matute. 2024. The Impact of AI Errors in a Human-in-the-Loop Process. *Cognitive Research: Principles and Implications* 9, 1 (2024), 1.

[3] Devdatta Akhawe and Adrienne Porter Felt. 2013. Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness. In *22nd USENIX Security Symposium (USENIX Security 13)*. USENIX Association, Washington, D.C., 257–272.

[4] Ahmed Aleroud and Lina Zhou. 2017. Phishing environments, techniques, and countermeasures: A survey. *Computers & Security* 68 (2017), 160–196.

[5] Kholoud Althobaiti, Adam DG Jenkins, and Kami Vaniea. 2021. A Case Study of Phishing Incident Response in an Educational Organization. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–32.

[6] Kholoud Althobaiti, Nicole Meng, and Kami Vaniea. 2021. I Don't Need an Expert! Making URL Phishing Features Human Comprehensible. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. ACM, Association for Computing Machinery, New York, NY, USA, Article 695, 17 pages. https://doi.org/10.1145/3411764.3445574

[7] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. arXiv:1606.06565 [cs.AI] https://arxiv.org/abs/1606.06565

[8] Eduardo Benavides, Walter Fuertes, Sandra Sanchez, and Manuel Sanchez. 2020. Classification of Phishing Attack Solutions by Employing Deep Learning Techniques: A Systematic Literature Review. 152 (2020), 51–64.

[9] Benjamin Berens, Katerina Dimitrova, Mattia Mossano, and Melanie Volkamer. 2022. Phishing awareness and education–When to best remind. In *Workshop on Usable Security and Privacy (USEC)*. Network and Distributed System Security (NDSS) Symposium.

[10] Benjamin Maximilian Berens, Florian Schaub, Mattia Mossano, and Melanie Volkamer. 2024. Better Together: The Interplay Between a Phishing Awareness Video and a Link-centric Phishing Support Tool. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–60.

[11] Panagiotis Bountakas, Konstantinos Koutroumpouchos, and Christos Xenakis. 2021. A Comparison of Natural Language Processing and Machine Learning Methods for Phishing Email Detection. In *Proceedings of the 16th International Conference on Availability, Reliability and Security*. Association for Computing Machinery, New York, NY, USA, 1–12.

[12] Cristian Bravo-Lillo, Saranga Komanduri, Lorrie Faith Cranor, Robert W Reeder, Manya Sleeper, Julie Downs, and Stuart Schechter. 2013. Your attention please: Designing security-decision UIs to make genuine risks harder to ignore. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*. Association for Computing Machinery, New York, NY, USA, 1–12.

[13] John Brooke. 1996. SUS: A 'Quick and Dirty' Usability Scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[14] Pavlo Burda, Luca Allodi, Alexander Serebrenik, and Nicola Zannone. 2024. 'Protect and Fight Back': A Case Study on User Motivations to Report Phishing Emails. In *Proceedings of the 2024 European Symposium on Usable Security*. Association for Computing Machinery, New York, NY, USA, 30–43.

[15] Pavlo Burda, Luca Allodi, and Nicola Zannone. 2020. Don't Forget the Human: A Crowdsourced Approach to Automate Response and Containment Against Spear Phishing Attacks. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE Computer Society, Los Alamitos, CA, USA, 471–476.

[16] Jason W Burton, Mari-Klara Stein, and Tina Blegind Jensen. 2020. A systematic review of algorithm aversion in augmented decision making. *Journal of behavioral decision making* 33, 2 (2020), 220–239.

[17] Marcus Butavicius, Ronnie Taib, and Simon J Han. 2022. Why people keep falling for phishing scams: The effects of time pressure and deception cues on the detection of phishing emails. *Computers & Security* 123 (2022), 102937.

[18] Jie Cai, Jiawei Luo, Shulin Wang, and Sheng Yang. 2018. Feature selection in machine learning: A new perspective. *Neurocomputing* 300 (2018), 70–79.

[19] Gamze Canova, Melanie Volkamer, Clemens Bergmann, and Benjamin Reinheimer. 2015. NoPhish App Evaluation: Lab and Retention Study. In *NDSS Workshop on Usable Security*. Internet Society, San Diego, CA, USA.

[20] The National Cyber Security Centre. 2018. Phishing attacks: defending your organisation. https://www.ncsc.gov.uk/guidance/phishing [Accessed 26-10-2023].

[21] Xiaowei Chen, Margault Sacré, Gabriele Lenzini, Samuel Greiff, Verena Distler, and Anastasia Sergeeva. 2024. The Effects of Group Discussion and Role-playing Training on Self-efficacy, Support-seeking, and Reporting Phishing Emails: Evidence from a Mixed-design Experiment. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–21.

[22] Leah Chong, Guanglu Zhang, Kosa Goucher-Lambert, Kenneth Kotovsky, and Jonathan Cagan. 2022. Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior* 127 (2022), 107018.

[23] WJ Conover. 1999. *Practical nonparametric statistics*. John Wiley & Sons, Inc.

[24] Verena Distler. 2023. The Influence of Context on Response to Spear-Phishing Attacks: an In-Situ Deception Study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–18.

[25] Nguyet Quang Do, Ali Selamat, Ondrej Krejcar, Enrique Herrera-Viedma, and Hamido Fujita. 2022. Deep Learning for Phishing Detection: Taxonomy, Current Challenges and Future Directions. *Ieee Access* 10 (2022), 36429–36463.

[26] Anne-Kee Doing, Eduardo Bárbaro, Frank van der Roest, Pieter van Gelder, Yury Zhauniarovich, and Simon Parkin. 2024. An Analysis of Phishing Reporting Activity in a Bank. In *Proceedings of the 2024 European Symposium on Usable Security*. Association for Computing Machinery, New York, NY, USA, 44–57.

[27] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. In *Proceedings of the 2017 chi conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 278–288.

[28] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. 2008. You've Been Warned: An Empirical Study of the Effectiveness of Web Browser Phishing Warnings. In *Proceedings of the SIGCHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1065–1074.

[29] Serge Egelman and Eyal Peer. 2015. The myth of the average user: Improving privacy and security systems through individualization. In *Proceedings of the 2015 New Security Paradigms Workshop*. Association for Computing Machinery, New York, NY, USA, 16–28.

[30] Serge Egelman and Eyal Peer. 2015. Scaling the Security Wall: Developing a Security Behavior Intentions Scale (SEBIS). In *Proceedings of the 33rd conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 2873–2882.

[31] Egress. 2024. Phishing Threat Trends Report. https://pages.egress.com/whitepaper-phishing-trends-threat-report-04-24.html Accessed: 2024-12-05.

[32] Egress Software Technologies Ltd. 2024. 2024 Email Security Risk Report. https://pages.egress.com/whitepaper-email-risk-report-01-24.html#download Accessed: 2024-07-19.

[33] Rasha Salah El-Din. 2012. To Deceive or Not to Deceive! Ethical Questions in Phishing Research. In *Electronic Workshops in Computing*. BCS Learning & Development.

[34] Adrienne Porter Felt, Alex Ainslie, Robert W Reeder, Sunny Consolvo, Somas Thyagaraja, Alan Bettes, Helen Harris, and Jeff Grimes. 2015. Improving SSL warnings: Comprehension and adherence. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 2893–2902.

[35] Ian Fette, Norman Sadeh, and Anthony Tomasic. 2007. Learning to Detect Phishing Emails. In *Proceedings of the 16th international conference on World Wide Web*. Association for Computing Machinery, New York, NY, USA, 649–656.

[36] Anjuli Franz, Verena Zimmermann, Gregor Albrecht, Katrin Hartwig, Christian Reuter, Alexander Benlian, and Joachim Vogt. 2021. SoK: Still Plenty of Phish in the Sea — A Taxonomy of User-Oriented Phishing Interventions and Avenues for Future Research. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*. USENIX Association, Boston, MA, 339–358.

[37] Gene V Glass. 1966. Testing Homogeneity of Variances. *American Educational Research Journal* 3, 3 (1966), 187–190.

[38] Daniel Jampen, Gürkan Gür, Thomas Sutter, and Bernhard Tellenbach. 2020. Don't click: Towards an effective anti-phishing training. A comparative literature review. *Human-centric Computing and Information Sciences* 10, 1 (2020), 33.

[39] Asangi Jayatilaka, Nalin Asanka Gamagedara Arachchilage, and Muhammad Ali Babar. 2021. Falling for Phishing: An Empirical Investigation into People's Email Response Behaviors. *arXiv preprint arXiv:2108.04766* (2021).

[40] Asangi Jayatilaka, Nalin Asanka Gamagedara Arachchilage, and Muhammad Ali Babar. 2024. Why People Still Fall for Phishing Emails: An Empirical Investigation into How Users Make Email Response Decisions. In *Symposium on Usable Security and Privacy (USEC)*. NDSS Symposium. https://www.ndss-symposium.org/wp-content/uploads/usec2024-72-paper.pdf

[41] Adam Jenkins, Nadin Kokciyan, and Kami E Vaniea. 2022. Phished: Automated Contextual Feedback for Reported Phishing. In *18th Symposium on Usable Privacy and Security [Poster Session]*. USENIX Security.

[42] Matthew L. Jensen, Alexandra Durcikova, and Ryan T. Wright. January 4-7, 2017. Combating Phishing Attacks: A Knowledge Management Approach. In *50th Hawaii International Conference on System Sciences, HICSS*. ScholarSpace / AIS Electronic Library (AISeL), Hilton Waikoloa Village, Hawaii, USA, 1–10. http://hdl.handle.net/10125/41681

[43] Matthew L Jensen, Ryan T Wright, Alexandra Durcikova, and Shamya Karumbaiah. 2022. Improving phishing reporting using security gamification. *Journal of Management Information Systems* 39, 3 (2022), 793–823.

[44] Amir Kashapov, Tingmin Wu, Sharif Abuadbba, and Carsten Rudolph. 2022. Email summarization to assist users in phishing identification. In *Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security*. Association for Computing Machinery, New York, NY, USA, 1234–1236.

[45] Leon Kersten, Pavlo Burda, Luca Allodi, and Nicola Zannone. 2022. Investigating the effect of phishing believability on phishing reporting. In *2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 117–128.

[46] Mahmoud Khonji, Youssef Iraqi, and Andrew Jones. 2013. Phishing Detection: A Literature Survey. *IEEE Communications Surveys & Tutorials* 15, 4 (2013), 2091–2121.

[47] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14.

[48] Nadin Kokciyan and Pınar Yolum. 2022. Taking Situation-Based Privacy Decisions: Privacy Assistants Working with Humans. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. International Joint Conferences on Artificial Intelligence Organization, 703–709.

[49] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. 2010. Teaching Johnny not to fall for phish. *ACM Transactions on Internet Technology (TOIT)* 10, 2 (2010), 1–31.

[50] Youngsun Kwak, Seyoung Lee, Amanda Damiano, and Arun Vishwanath. 2020. Why do users not report spear phishing emails? *Telematics and Informatics* 48 (2020), 101343.

[51] Daniele Lain, Kari Kostiainen, and Srdjan Čapkun. 2022. Phishing in organizations: Findings from a large-scale and long-term study. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 842–859.

[52] Nhien-An Le-Khac and Tahar Kechadi. 2015. Security threats of url shortening: a users perspective. (2015).

[53] Deyi Li and Yi Du. 2017. *Artificial intelligence with uncertainty*. CRC press.

[54] Richard Lowry. 2008. One way ANOVA–independent samples. *Vassar. edu* (2008).

[55] Federico Maggi, Alessandro Frossi, Stefano Zanero, Gianluca Stringhini, Brett Stone-Gross, Christopher Kruegel, and Giovanni Vigna. 2013. Two years of short urls internet measurement: security threats and countermeasures. In *Proceedings of the 22nd International Conference on World Wide Web*. Association for Computing Machinery, New York, NY, USA, 861–872.

[56] Ioana Andreea Marin, Pavlo Burda, Nicola Zannone, and Luca Allodi. 2023. The Influence of Human Factors on the Intention to Report Phishing Emails. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 620, 18 pages. https://doi.org/10.1145/3544548.3580985

[57] John Marsden, Zachary Albrecht, Paula Berggren, Jessica Halbert, Kyle Lemons, Anthony Moncivais, and Matthew Thompson. 2020. Facts and stories in phishing training: A replication and extension. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–6.

[58] Nina Marshall, Daniel Sturman, and Jaime C Auton. 2023. Exploring the evidence for email phishing training: A scoping review. *Computers & Security* 139 (2023), 103695.

[59] Mattia Mossano, Kami Vaniea, Lukas Aldag, Reyhan Düzgün, Peter Mayer, and Melanie Volkamer. 2020. Analysis of publicly available anti-phishing webpages: contradicting information, lack of concrete advice and very narrow attack vector. In *Proceedings of the 5th IEEE European Workshop on Usable Security (EuroUSEC)*. IEEE. https://doi.org/10.1109/EuroSPW51379.2020.00026

[60] National Institute of Standards and Technology. 2024. The NIST Cybersecurity Framework (CSF) 2.0. https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf

[61] Jose Nazario. 2005. Nazario Phishing Corpus. https://monkey.org/~jose/phishing/, (accessed on 16 Feb 2023).

[62] Q.H. Nguyen, T. Wu, V. Nguyen, X. Yuan, and J. Xue. 2024. Utilizing Large Language Models with Human Feedback Integration for Generating Dedicated Warning for Phishing Emails. In *Proceedings of the 2nd ACM Conference on Information Technology and Security*. ACM, Association for Computing Machinery, New York, NY, USA, 35–46.

[63] Kathryn Parsons, Marcus Butavicius, Malcolm Pattinson, Dragana Calic, Agata Mccormac, and Cate Jerram. 2016. Do Users Focus on the Correct Cues to Differentiate Between Phishing and Genuine Emails? *arXiv preprint arXiv:1605.04717* (2016).

[64] Justin Petelka, Yixin Zou, and Florian Schaub. 2019. Put Your Warning Where Your Link Is: Improving and Evaluating Email Phishing Warnings. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1–15.

[65] Nikolas Pilavakis, Adam Jenkins, Nadin Kokciyan, and Kami Vaniea. 2023. "I didn't click": What users say when reporting phishing. In *Proceedings of the Symposium on Usable Privacy and Security (USEC'23)*. The Internet Society, Reston, VA, 1–13. https://doi.org/10.14722/usec.2023.233129

[66] Amy Rechkemmer and Ming Yin. 2022. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. In *Proceedings of the 2022 chi conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1–14.

[67] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. 2016. How I Learned to be Secure: a Census-Representative Survey of Security Advice Sources and Behavior. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. Association for Computing Machinery, New York, NY, USA, 666–677.

[68] Benjamin Reinheimer, Lukas Aldag, Peter Mayer, Mattia Mossano, Reyhan Duezguen, Bettina Lofthouse, Tatiana Von Landesberger, and Melanie Volkamer. 2020. An investigation of phishing awareness and education over time: When and how to best remind users. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*. USENIX Association, Boston, MA, 259–284.

[69] Tarini Saka, Rachiyta Jain, Kami Vaniea, and Nadin Kökciyan. 2024. Phishing Codebook: A Structured Framework for the Characterization of Phishing Emails. arXiv:2408.08967 [cs.CR] https://arxiv.org/abs/2408.08967

[70] Tarini Saka, Kami Vaniea, and Nadin Kökciyan. 2022. Context-Based Clustering to Mitigate Phishing Attacks. In *Proceedings of the 15th ACM Workshop on Artificial Intelligence and Security*. Association for Computing Machinery, New York, NY, USA, 115–126.

[71] Tarini Saka, Kami Vaniea, and Nadin Kökciyan. 2024. PhishCoder: Efficient Extraction of Contextual Information from Phishing Emails. In *Proceedings of the Workshop on Security and Artificial Intelligence (SECAI 2024)*. https://drive.google.com/file/d/1441TcE6Wkc312wTTq5Y4haIWwnhKieBg/view

[72] Orvila Sarker, Sherif Haggag, Asangi Jayatilaka, and Chelsea Liu. 2023. Personalized Guidelines for Design, Implementation and Evaluation of Anti-phishing Interventions. In *2023 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*. IEEE, 1–12.

[73] Dawn M Sarno and Mark B Neider. 2022. So Many Phish, So Little Time: Exploring Email Task Factors and Phishing Susceptibility. *Human Factors* 64, 8 (2022), 1379–1403.

[74] Lorin Schöni, Victor Carles, Martin Strohmeier, Peter Mayer, and Verena Zimmermann. 2024. You Know What?-Evaluation of a Personalised Phishing Training Based on Users' Phishing Knowledge and Detection Skills. In *Proceedings of the 2024 European Symposium on Usable Security*. Association for Computing Machinery, New York, NY, USA, 1–14.

[75] Samuel Sanford Shapiro and Martin B Wilk. 1965. An Analysis of Variance test for Normality (complete samples). *Biometrika* 52, 3-4 (1965), 591–611.

[76] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. 2007. Anti-Phishing Phil: The Design and Evaluation of a Game That Teaches People Not to Fall for Phish. In *Proceedings of the 3rd symposium on Usable privacy and security*. Association for Computing Machinery, New York, NY, USA, 88–99.

[77] Ivan Skula and Michal Kvet. 2024. URL and Domain Obfuscation Techniques-Prevalence and Trends Observed on Phishing Data. In *2024 IEEE 22nd World Symposium on Applied Machine Intelligence and Informatics (SAMI)*. IEEE, 000283–000290.

[78] Nathalie Stembert, Arne Padmos, Mortaza S Bargh, Sunil Choenni, and Frans Jansen. 2015. A Study of Preventing Email (Spear) Phishing by Enabling Human Intelligence. In *2015 European intelligence and security informatics conference*. IEEE, 113–120.

[79] Michelle Steves, Kristen Greene, and Mary Theofanos. 2020. Categorizing human phishing difficulty: a Phish Scale. *Journal of Cybersecurity* 6, 1 (2020), tyaa009.

[80] Mark Steyvers and Aakriti Kumar. 2023. Three challenges for AI-assisted decision-making. *Perspectives on Psychological Science* 19, 5 (2023), 17456916231181102.

[81] Sean Strain. 2022. *Automatically Generating Contextualised Responses to Phishing Reports*. Master of Informatics thesis (Part 1). University of Edinburgh. Available at https://tulipslab.org/projects/21-22/Sean-Strain.pdf.

[82] Sean Strain. 2023. *Evaluating the Impact of an Automated Email Analysis System on Phishing Susceptibility*. Master of Informatics thesis (Part 2). University of Edinburgh. Available at https://tulipslab.org/projects/22-23/strain-thesis.pdf.

[83] Jenny Tang, Eleanor Birrell, and Ada Lerner. 2022. Replication: How Well Do My Results Generalize Now? The External Validity of Online Privacy and Security Surveys. In *Eighteenth symposium on usable privacy and security (SOUPS 2022)*. USENIX Association, Boston, MA, 367–385.

[84] Tessian. 2022. Must-Know Phishing Statistics: Updated 2022. https://www.tessian.com/blog/phishing-statistics-2020/. https://www.tessian.com/blog/phishing-statistics-2020/ Accessed: 2024-09-09.

[85] George Thomopoulos, Dimitrios Lyras, and Christos Fidas. 2023. Methodologies and Ethical Considerations in Phishing Research: A Comprehensive Review. In *Proceedings of the 2nd International Conference of the ACM Greek SIGCHI Chapter*. Association for Computing Machinery, New York, NY, USA, 1–10.

[86] Amber van der Heijden and Luca Allodi. 2019. Cognitive Triaging of Phishing Attacks. In *28th USENIX Security Symposium (USENIX Security 19)*. USENIX Association, Santa Clara, CA, 1309–1326. https://www.usenix.org/conference/usenixsecurity19/presentation/van-der-heijden

[87] Melanie Volkamer, Karen Renaud, Benjamin Reinheimer, and Alexandra Kunz. 2017. User experiences of TORPEDO: TOoltip-poweRed Phishing Email DetectiOn. *Computers & Security* 71 (2017), 100–113.

[88] Jaclyn Wainer, Laura Dabbish, and Robert Kraut. 2011. Should I Open this Email?: Inbox-Level Cues, Curiosity and Attention to Email. In *Proceedings of the SIGCHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 3439–3448.

[89] Xinru Wang, Zhuoran Lu, and Ming Yin. 2022. Will you accept the ai recommendation? predicting human behavior in ai-assisted decision making. In *Proceedings of the ACM web conference 2022*. Association for Computing Machinery, New York, NY, USA, 1697–1708.

[90] Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, New York, NY, USA, 318–328.

[91] Rick Wash. 2020. How Experts Detect Phishing Scam Emails. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–28.

[92] Rick Wash and Molly M Cooper. 2018. Who Provides Phishing Training?: Facts, Stories, and People Like Me. In *Proceedings of the 2018 chi conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1–12.

[93] Rick Wash, Norbert Nthala, and Emilee Rader. 2021. Knowledge and Capabilities that Non-Expert Users Bring to Phishing Detection. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*. USENIX Association, Boston, MA, 377–396.

[94] Aiping Xiong, Robert W Proctor, Weining Yang, and Ninghui Li. 2019. Embedding Training Within Warnings Improves Skills of Identifying Phishing Webpages. *Human factors* 61, 4 (2019), 577–595.

[95] Weining Yang, Jing Chen, Aiping Xiong, Robert W Proctor, and Ninghui Li. 2015. Effectiveness of a Phishing Warning in Field Settings. In *Proceedings of the 2015 Symposium and Bootcamp on the Science of Security*. Association for Computing Machinery, New York, NY, USA, 1–2.

[96] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852

[97] Sijie Zhuo, Robert Biddle, Yun Sing Koh, Danielle Lottridge, and Giovanni Russello. 2023. SoK: Human-centered Phishing Susceptibility. *ACM Transactions on Privacy and Security* 26, 3 (2023), 1–27.

APPENDIX

Email Response Patterns for Inaccurate Reports

R2_1: The phishing email was purportedly from PayPal. The sender's address appeared as 'paypal.co.uk', while the link domain was 'paypal.com'. The primary tactic of the scam involved a fake link that led nowhere, forcing recipients to call a phone number provided in the email. The `Perfect` group participants were told the email was a scam and warned about non-ASCII characters, leading 81.88% to correctly identify it as phishing. Most cited the scam classification in their reasoning, while those who misclassified it cited the link destination. The `Realistic` group participants were told the email was safe, with the sender address and link domain provided as proof.

Only 33.13% identified the email as a scam, often citing the sender domain. The remaining who classified it as safe often cited the overall report for their decision. *Wrong classification with evidence such as spoofed domains, which are quite common, can hugely impact user response.*

R2_2: The phishing email in this case pretended to be from the *<Blinded>* university's support team, claiming the user's password was expiring, and hence was chosen to be contextually relevant. Both the sender domain and link destination were flagged as external. Email account-based phishing schemes are increasingly common due to their relevance to most users [69]. In the `Perfect` group, 95.63% correctly identified the email as phishing, with most pointing to the sender domain as the key indicator. In the `Realistic` group, participants were told the email was a scam but were shown a legitimate university domain as the link destination, but 96.99% still identified it as phishing, again citing the sender address. Those who misclassified the email most often relied on the overall report. *Wrongly parsed and shown evidence had no impact on the user performance if the high-level classification is correct.*

R2_4: The phishing email claimed to be from the Driver and Vehicle Licensing Agency (DVLA), offering recipients a tax refund. Both the sender domain and link destination were flagged as external. In the `Perfect` group, 97.50% of participants correctly identified the email as phishing, citing the sender domain as the key indicator, having been shown the real sender and link domains marked as suspicious. The email included a fake link with the domain 'service.gov.uk' to mislead users. In the `Realistic` group, were told the email was safe and shown the legitimate domain. This led to only 77.11% recognizing it as phishing, again citing the sender address, while those who misclassified it often pointed to the link destination as their reasoning. *Wrong classification with false evidence (which users can tell is wrongly parsed) can have a small impact.*

R2_8: This was a benign email from Amazon, containing a link for recipients to track their package, which was out for delivery. In the `Perfect` group, 96.25% of participants correctly identified the email as benign, with most citing the overall report as their reason. The `Realistic` group's report classified the email as safe, with a lower percentage of 55% compared to the perfect group (95%). Despite this, 90.96% of participants still correctly identified it as phishing. Those who misclassified it as a scam pointed to the high-level classification or mentioned they did not rely on the report in making their choice. However, participant feedback mentions that this affected their confidence; Eg: *"There's nothing I can really think of apart from being more certain that the email was either legit or a scam as some percentages were say 55% which didn't feel me with much confidence." Low classification percentages, which are common as ML is probabilistic, can impact user confidence in the system.*

R2_10: The was a benign email from Standard Life Investments about a summer internship, with the sender domain coming from within the university and the link being a shortened URL (bit.ly). This email had the lowest performance across all report conditions. In the `Perfect` group, participants were informed that the email was 95% likely to be safe, and 78.13% correctly identified it as such, most commonly citing the sender domain as their reason. Those who misclassified it as phishing often pointed to the unknown link destination. In the `Realistic` group, the email was classified as a financial scam with a 65% likelihood. Only 22.2% of participants identified it as safe, often citing the sender domain or disregarding the report. However, 77.8% followed the report and flagged the email as phishing based on the link destination. *Commonly used obfuscation techniques, like shortened URLs, are suspicious to users. Along with the wrong classification, this can have a huge impact on users.*

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009