Enhancing Phishing Detection: An Eye-Tracking Study on User Interaction and Oversights in Phishing Emails

Meret Kristen, Fabian Engl, Jürgen Mottok
Software Engineering Laboratory for Safe and Secure Systems
OTH Regensburg
Regensburg, Germany
email: {meret.kristen | fabian.engl | juergen.mottok}@oth-regensburg.de

***Abstract*—Phishing remains a significant threat to organizational security, necessitating effective countermeasures. This paper presents findings from an in-depth eye-tracking study with 103 participants, evaluating the effectiveness of phishing awareness tools and trainings. The study examines how a phishing awareness system influences user behavior, efficiency, and the ability to identify phishing attempts. By analyzing eye movements, the study reveals real-time interactions and oversights, providing insights into the decision-making process. Results indicate that while the system improves the efficiency of users already proficient in phishing detection, it does not universally enhance recognition rates. Notably, participants using the tool spent significantly less time looking at attachment-related phishing markers, indicating partial efficiency improvements. Since phishing attempts containing suspicious attachments were successful in 19% of cases, as compared to an overall phishing success rate of 15%, the phishing awareness tool is particularly useful here. A usability evaluation revealed that users reporting a higher perceived usability score profited more from the help of the tool. Additionally, no improvement in phishing detection rates was observed in users who had completed prior IT-security training, highlighting the necessity for a paradigm shift in phishing training to adequately prepare users for phishing attempts.**

Keywords-Phishing; Security Awareness; Eye-Tracking; IT-Security; Usability and UX.

I. Introduction

In information security, various components must work together to form a robust and secure system, with one of the biggest vulnerabilities in this chain being the end user [1]. Regardless of the amount of time and money an organization invests in cybersecurity, the risk of an incident increases significantly if an end user clicks on a compromised link or opens a hazardous attachment. For this reason, the ISO 27001 clause 7.2.2 states 'All employees of the organization and, where relevant, contractors should receive appropriate awareness education and training and regular updates in organizational policies and procedures, as relevant for their job function' [2]. But how effective are these trainings and procedures? Can tools help the end user to distinguish between ordinary emails and phishing attempts? And if they fall for a phishing attack, what relevant information did they ignore?

Publications, such as [3], show that, in contrast to existing works on phishing training, such as [4], voluntary contextual phishing trainings can have the opposite effect, making employees even more susceptible to phishing attacks. While incorporating warnings into email software improves effectiveness, the extent of the warnings — whether short or detailed — does not significantly influence their effectiveness. However, it is unclear to what degree users used the supplied warnings or how it changes the user's behaviour when interacting with emails. Such interactions are difficult to measure as only considering the end results of phishing studies does not paint a clear picture of the subconscious intentions of users when analysing phishing emails.

It has been proven that technologies like eye-tracking enable the measurement of such interactions in real time, showing that eye movements are directly related to thought processes when users view specific information [5]. This is particularly helpful when trying to evaluate the usability and effectiveness of tools, such as phishing awareness software. By analysing the eye movements of participants when working with such tools a pool of further metrics can be measured revealing real time information about decision making, whether users observe every part of the email and which areas were overlooked when participants fail to recognise a phishing attempt. For these reasons this paper proposes a phishing email study based on eye-tracking data and analyses whether supportive security awareness tools can help users to identify phishing emails.

The paper is structured as follows: first, a systematic literature review in Section II identifies the current state of knowledge and research gaps, followed by the formulation of research questions and corresponding hypotheses. This is followed by the study design in Section III, results in Section IV, usability results in Section V, limitations in Section VI and conclusion and future work in Section VII.

II. Literature Review

Having established the benefits of incorporating eye movements in phishing research, a systematic literature review based on the methods given by Kitchenham and Charters in [6] was performed. To establish a broad overview of the status quo of eye-tracking research in the field of security awareness and phishing emails lead to the following research question:

RQ1 What is the state of the art in eye-tracking research for detecting and analyzing user interaction with phishing emails?

As the study of phishing emails and security awareness is a critical part of engineering safe and secure systems, three of the main academic search engines in software engineering

were employed: IEEE Xplore, ACM Digital Library and Web of Science. To study the aforementioned research question, a search string was developed based on the methods given in [7]. The partial search string on eye-tracking should thus include the terms for method and device each in the two common spelling variants with and without hyphens (i.e. ”eye tracking”, ”eye-tracking”, ”eye tracker”, and ”eye-tracker”). The partial search string on phishing emails and security awareness was chosen to include only terms that are directly related to phishing emails ("phishing", "security awareness", "spam", "social engineering"). Since the term "email" always appears together with one of the search terms stated above in the context of phishing emails, it was not explicitly included in the search string.

(”eyetracking” OR ”eye-tracking” OR ”eyetracker” OR ”eye-tracker” OR "eye movement" OR "eye movements") AND ("phishing" OR "security awareness" OR "spam" OR "social engineering")

Due to the different search engines needing different input syntax, the actual search queries used differ slightly in syntax, but not in semantics. The search queries used are shown in Figure 1.

The search yielded three results in ACM, twelve in Web of Science and six in IEEE XPLORE as of February 2024, including one duplicate, giving 20 results in total.

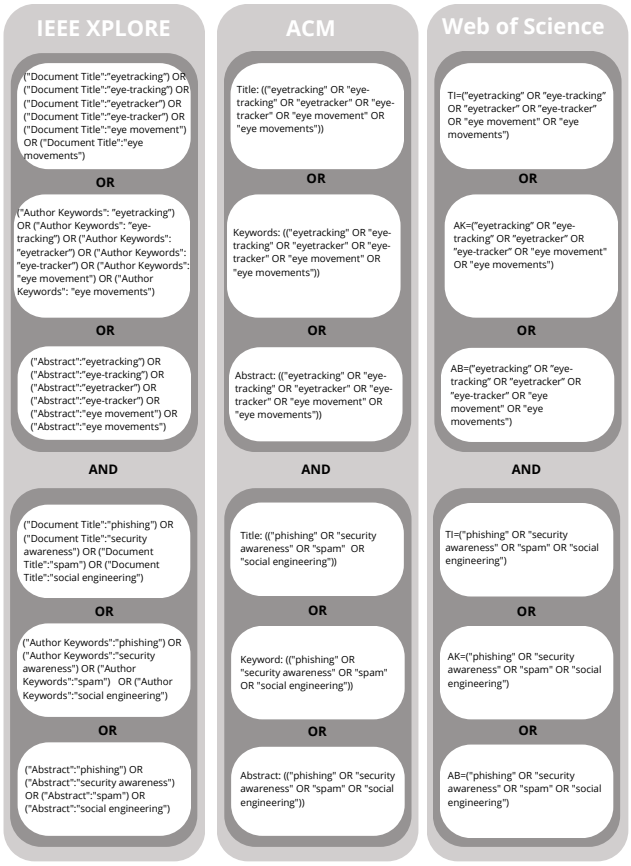

Figure 1. Search strings by data base.

A. Inclusion and Exclusion Criteria

Inclusion and exclusion criteria were defined as follows: Papers need to

1) study email phishing attempts, and
2) conduct eye-tracking studies or evaluate existing eye-tracking data and
3) be accessible with licenses held by OTH Regensburg or University of Regensburg.

After applying the inclusion and exclusion criteria, a total of eight papers and an additional four papers after backward and forward search remained. The found papers are described in the following list.

1) *ADVERT: An Adaptive and Data-Driven Attention Enhancement Mechanism for Phishing Prevention* [8]: This paper presents a study evaluating the effectiveness of generating adaptive visual aids in real-time to prevent user inattentiveness and reduce susceptibility to phishing attacks. The study was conducted with a sample size of 160 students and involved twelve emails.
2) *Evaluation of Contextual and Game-Based Training for Phishing Detection* [9] A study with 41 participants tasked with identifying phishing emails, divided into three groups: without prior training, with game-based training, and with Context-Based Micro-Training (CBMT). The research shows that both training methods can support users towards secure behavior and that CBMT does so to a higher degree than game-based training. In line with [3], the paper also shows that most participants were susceptible to phishing, even after training.
3) *Understanding Phishing Email Processing and Perceived Trustworthiness Through Eye Tracking* [10] In this pilot study, a group of 22 volunteers saw a sequence of emails that included or did not contain signs of phishing emails, all the while having their eye movements monitored. Despite the fact that the phishing signs demanded a higher attentional investment, the study demonstrates that less time was spent viewing them.
4) *Investigating Gaze Behavior in Phishing Email Identification* [11] A preliminary study including 28 students revealing that specialists perform better at identifying phishing emails and that experts and non-experts use different techniques for email examination.
5) *Perceiving and Using Genre by Form – An Eye-Tracking Study* [12] A study with 24 participants tasked with classifying emails into genres (calls for papers, newsletters, spam) demonstrated that genre analysis based on purpose and form is an effective method for identifying the characteristics of these emails. This paper is not specific to phishing emails.

6) *You have e-mail, what happens next? Tracking the eyes for genre* [13] A follow-up paper to the previous eye-tracking study by Clark, with further insight on how users classify emails into genres.
7) *Prediction of Phishing Susceptibility Based on a Combination of Static and Dynamic Features* [14] The user phishing susceptibility prediction model (DSM) presented in this paper is built on a combination of static and dynamic variables. A study involving 50 participants in eye-tracking was carried out to confirm that the model correctly predicts the behavior.
8) *Eyes wide open: The role of situational information security awareness for security-related behaviour* [15] Provides thorough literature research on empirical phishing research and conducts a study with 107 participants to examine how individual-level and system-level factors influence awareness. The findings highlight the significance of situational information security awareness and demonstrate that, whereas contextual relevance and misplaced salience in phishing emails reduce awareness, prior exposure to phishing and security warnings increases awareness.
9) Further papers on [8] with more detailed statistical analyses of the same study
10) *Where the User Does Look When Reading Phishing Mails - An Eye-Tracking Study* [16] A study with 25 participants that were shown emails and decided whether they were phishing. The findings indicate that two critical elements in identifying phishing emails are time and expertise.
11) *Email Reading Behavior-Informed Machine Learning Model to Predict Phishing Susceptibility* [17] A prototype tested with 25 participants to collect eye-tracking data in real time and notify users when they are on the brink of falling for phishing.
12) *Revealing the Hidden Effects of Phishing Emails: An Analysis of Eye and Mouse Movements in Email Sorting Tasks* [18], An online study with 39 participants using mouse movements and gaze patterns. The study shows that when interacting with phishing versus non-phishing emails, there are notable changes in mouse movements and eye gaze.

B. Results of the Literature Review

The literature review shows that while there are previous empirical studies on user interaction with phishing emails that analyze eye movements, the papers found either have a relatively small sample size, or study adaptive mechanisms meant to improve the users phishing recognition. A clear research gap in studying how participants use the provided tools and warnings and which phishing markers that should have raised suspicion were overlooked when users fall for a phishing attempt can be identified. These questions need to be studied in order to develop tools and strategies to prevent phishing attacks. Based on this literature review, the following research questions were developed:

RQ2 How does the use of an additional phishing awareness system influence the effectiveness of recognition of phishing emails?
RQ3 How does the use of an additional phishing awareness system influence the efficiency of the recognition of phishing emails?
RQ4 How does the existence of the phishing awareness system influence the amount of time spent looking at phishing markers?
RQ5 Which phishing markers of an email are most commonly overlooked when a user falls for a phishing attempt?

Based on these research questions, the following hypothesis were developed:

H1 Participants with the phishing awareness system will correctly identify a higher percentage of emails compared to the group not using the sidebar.
H2 Participants with the phishing awareness system need less time to classify the email.
H3 Participants with the phishing awareness system spend less time looking at the relevant phishing markers before making a decision.
H4 Participants with the phishing awareness system that recognise a phishing attempt spent less time proportionally looking at phishing markers compared to participants with the phishing awareness system that fall for a phishing attempt.

III. Study design

The study included 18 different stimuli: twelve phishing emails and six harmless control emails. All emails were real, with minor modifications made to obscure personal details. Furthermore, one email was translated from English to German to eliminate potential language barriers. The twelve phishing emails were further divided evenly into the following categories, to cover a wide spectrum of typical phishing emails:

1) containing a suspicious attachment
2) containing a link to an external website and an injunction to click on said link
3) containing an injunction to send money or items of value (e.g., gift cards, sensitive data)

Each of the three categories is split into two subgroups containing two emails each. This separation is based on the quality of the phishing email, which is measured by the amount of phishing markers within an email. Phishing markers are defined as elements that indicate phishing emails, such as spelling errors, cryptic text, misleading domains or suspicious attachments like $.exe$ or $.docm$. For this study, a well-made phishing email is defined as containing a maximum of two subtle phishing markers, such as a slightly altered domain name like $@spotfy.com$. In contrast, poorly written phishing emails are characterized by having more than two markers or very obvious signs, such as cryptic sender addresses. Due to the subjectivity of the interaction with the email, it has to be noted that these categories are not always precisely distinguished and may overlap. The control group also consists

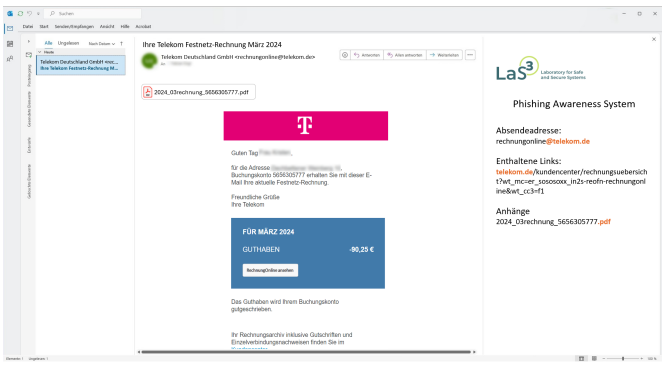

Figure 2. A phishing email with the Phishing Awareness Sidebar (PAS).

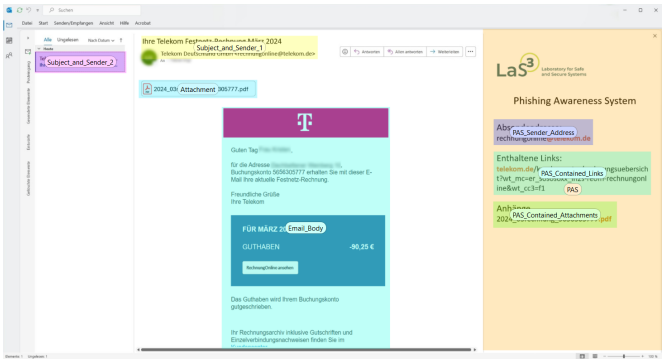

Figure 3. Defined AOIs on an email with the PAS.

of six emails, two for each of the categories names above. The emails were presented in random order.

A. Phishing Awareness System (PAS)

In order to study the research questions mentioned above, a prototype for a Phishing Awareness System (PAS) that is similar to ones already on the market was build. It was embedded into Microsoft Outlook (see Figure 2), as it is a commonly used email client in an office environment. The prototype was designed to help user identify the most common phishing markers by highlighting them. These markers include suspicious links, attachments and the address of the sender. Participants using the system were informed about the existence of the PAS and its functions beforehand.

An in-between-subjects design was used in this study, where half of the participants were provided with an Outlook environment that included the PAS, while the other half used a standard Outlook environment. Group assignment was done randomly to ensure unbiased distribution.

B. Participants

A total of 120 participants were recruited from various local small and medium enterprises, as well as public sector organizations, to ensure a representative real-world dataset. Eleven participants chose not to answer the questionnaire and were subsequently excluded from the dataset. An additional six participants did not meet the calibration and validation requirement of 0.75°, primarily due to extreme visual impairments. Despite this, these six participants still wished to participate in the study for personal interest but were informed that their data would not be considered in the final study. Of the remaining 103 participants, 51 performed the study with the PAS and 52 without. The mean age was 35 for the PAS group and 34 for the group without PAS. In the PAS group, 69% of the participants were male and 31% were female, whereas in the group without the sidebar, 58% of the participants were male and 42% were female. In both groups, over 90% of users (92% with the PAS and 90% without) reported knowing what phishing emails can look like and being able to identify suspicious features. Additionally, 57% of participants in the PAS group and 71% in the group without the sidebar indicated that they receive phishing emails daily or several times per week. In the PAS group, only 49% of participants had previously participated in phishing training, compared to 71% in the no-sidebar group. While the age distribution and prior knowledge were nearly identical across both groups, there were significant differences in gender distribution, prior exposure to phishing, and experience with phishing training, which can influence the final results.

C. Eye-Tracking Setup and Data Collection

Nine mobile Tobii Pro Fusion eye-trackers running at 250 Hz were used for data collection, attached to modular 21-inch screens and each equipped with dedicated laptops. Participants were calibrated with a 65 cm distance to the eye-tracker and asked to sit still during the recording. A nine point calibration and four point validation was chosen to ensure optimal accuracy. To further ensure an accurate dataset, a quality threshold for calibration and validation was set to 0.75°.

The recording locations varied, as the study was conducted across a range of companies. In each location, the eye-trackers were set up in dedicated rooms, with blinds closed whenever possible to minimize direct natural light interference. Before the study, participants were informed about the procedures and asked to sign a consent form approved by the Joint Ethics Committee of the Bavarian Universities (GEHBa). Participation was voluntary and each participant was assigned an anonymous identifier. No additional phishing warnings or trainings were provided, as participants were aware they were participating in phishing research, which could lead to priming effects.

After being briefed on the study, participants were paired with a researcher and seated in front of a Tobii Pro Fusion eye-tracker equipped with a keyboard and a mouse. The eye-tracker was calibrated to each participant before the session began. Participants were initially shown two slides containing instructions with the group with the PAS receiving an additional slide explaining the sidebar's purpose. Participants could start the study at their own pace and had no maximum time to finish. Before each email, a centering cross appeared on the left side of the screen to ensure that participants started

viewing the stimulus from a neutral point. If they identified an email as phishing, they were instructed to press the "S" key; if they believed it was not phishing, they were to press the "Right" key. The two keys were purposely selected as the distance between them minimized the risk for accidental presses. Pressing either key would proceed to the next email stimulus.

After the eye-tracking experiment, each participant was given a questionnaire collecting demographic data, prior knowledge of IT-security topics and their familiarity with the companies mentioned in the emails. The participants with the PAS were also asked to rate the tool using the short version of the User Experience Questionnaire (UEQ-S) and the System Usability Scale (SUS).

D. Areas of Interest (AOIs)

Areas of Interest (AOIs) are regions predefined by the researchers that hold particular significance for the research subject. They represent various metrics, such as fixations occurring within a specific area [19]. In this study, the AOIs correspond to phishing markers present in emails and were drawn to match the areas containing these phishing markers, including the sender's email address, the email's subject line, the main body of the email, and any attachments. An illustration of these AOIs can be seen in Figure 3.

IV. RESULTS

For the first hypothesis H1, it was found that participants who used the sidebar generally did not perform better in the emails sorting task.

Figures 4 and 5 show that the sidebar group was overall less effective in correctly identifying emails, as well as less effective in identifying phishing emails. A Shapiro-Wilk test [20] showed that the samples are non-normally distributed, proving the need for non-parametric tests to be performed. A Mann-Whitney U-test with approximated p-value on the number of correctly identified emails in both groups fails to show a difference in distribution between the groups: at $\alpha = 0.05$ it results in $z = 1216.00$, $p = .460$, $r = .07$. A similar result holds for the number of correctly identified phishing emails; here, a Mann-Whitney U-test delivers $z = 1201.50$, $p = .396$, $r = .08$, showing no significant difference at $\alpha = 0.05$ between the two groups. This leads to having to reject H1, showing that there is no significant difference between the group with the sidebar and the group without when it comes to correctly identifying emails.

To test the second hypothesis H2, a Shapiro-Wilk test on the dependent variable 'total time' showed no normal distribution within both groups. This means again that a non-parametric test should be used. A Mann-Whitney U-test yields $z = 1518.00$, $p = .205$, $r = .12$ and thus revealing no significant difference between groups at $\alpha = 0.05$ with small effect. This can also be seen in Figure 6. Testing instead only the time spent to sort phishing emails gives a similar result: no significant difference between the group with PAS and the group without. Looking only at the time spent on emails of specific types (with attachment, with links or with an injunction to send money) also showed no significant difference between the two groups. Neither could a difference be found when looking only at good phishing emails, bad phishing emails or only the control group. Hypothesis 2 thus also has to be rejected.

Hypothesis H3 is the first hypothesis based on the col-

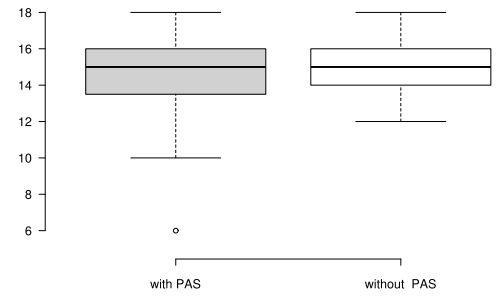

Figure 4. Total number of correctly identified emails with and without sidebar tool.

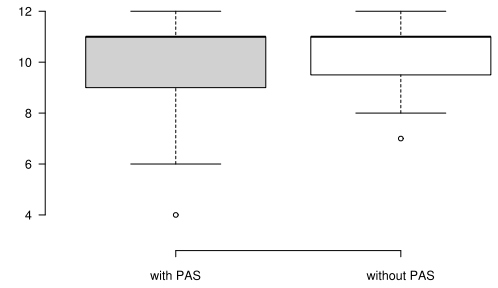

Figure 5. Total number of correctly identified phishing emails with and without sidebar tool.

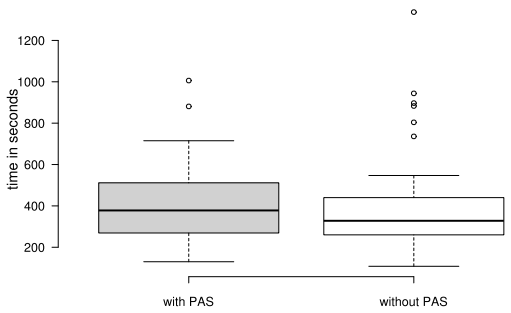

Figure 6. Total time spent on the email sorting task for groups with and without sidebar tool.

TABLE I. MANN-WHITNEY U-TEST RESULTS FOR AOI HITS BETWEEN THE GROUP WITH PAS AND THE GROUP WITHOUT PAS.

	z	p	r
Subject and Sender	1546.00	.147	.14
Attachments	1786.00	.002	.3
Email Body	1290.00	.812	.02

lected eye-tracking data, focusing on three different AOIs: fixations on the main body of the email, the attachments, and the sender's address along with the email's subject line. In Microsoft Outlook, the sender's address and subject line are displayed twice (once on the left side and once above the email) which have been consolidated into a single AOI for this analysis. For the group using the PAS, fixations on relevant phishing markers in both the tool and within Microsoft Outlook were combined. As before a Shapiro-Wilk test on the dependent variable 'AOI hits' showed no normal distribution within both groups. Hence, Mann-Whitney U-tests were applied to assess the number of AOI hits in both groups, as shown in Table I.

Both the number of AOI hits within the subject and sender information, as well as the email body, show no significant differences between the two groups at $\alpha = 0.05$ with small to neglectable effect sizes of .14 and .02, respectively. These small effect sizes indicate minimal differences between the groups. However, for the attachments, the p-value of .002 is well below the $\alpha = 0.05$ threshold, indicating a statistically significant difference in the number of AOI hits and, consequently, the amount of time spent looking at the attachments between the groups. The effect size r of .30 suggests a small to medium effect, indicating that the difference is not only statistically significant but also has moderate practical significance. As shown in Figure 7, the group with the PAS has significantly less AOI hits on the attachment ($M = 1435.58$) compared to the group without the assisting sidebar ($M = 2527.78$). In terms of time, the PAS group spent an average of 5.74 seconds looking at the attachments, compared to 10.11 seconds for the group without the tool.

Based on these findings, Hypothesis 3 can only be partially accepted. While there is no statistically significant difference between the two groups in viewing phishing markers in the email body or sender's address and subject line, there is a significant difference in AOI hits and therefore viewing time for attachments. This indicates that the presence of the PAS significantly reduces the time spent analyzing attachment types. These results underscore the relevance of eye-tracking technology in capturing not only easily measurable metrics, such as completion time which stayed the same between the two groups, but also subconscious interactions and relevant regions revealed by users' eye movements. This data is particularly valuable for understanding how users interact with phishing emails and for identifying which phishing markers attract the most attention or are overlooked.

The observation that participants using the PAS showed no significant difference in the overall time spent classifying

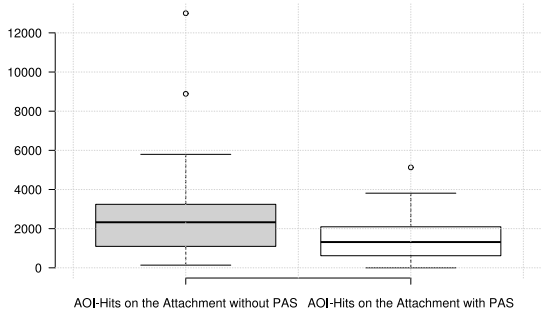

Figure 7. AOI hits on the attachment with and without the PAS.

emails with attachments, despite spending nearly half as much time looking at the attachments and their types, may be due to several explanations. For instance participants might not fully trust the tool and therefore seek to verify their decisions by examining additional phishing markers. Alternatively, the time difference might be attributed to the need to process the additional information provided by the tool.

To answer RQ5, it is necessary to look separately at the group with PAS and the group without. The reason for this is that certain phishing markers, such as sender address, attachments and contained links, are repeated in the PAS and thus participants might divide their attention between the phishing markers in the email and in the PAS. Another reason is that the PAS, being a new tool that participants have not used before, can attract attention from participants. In order not to skew the results, the evaluation was performed separately for the two groups.

For the group without the tool, looking at all 12 phishing emails, it was studied whether participants that classified the email correctly spent less time looking at the phishing markers contained in the AOIs "subject and sender", "email body" and "attachments" than participants that did not classify the email correctly. There was no statistically significant difference in the AOI hits for "subject and sender" and "attachments" found between the group that sorted the emails correctly and the group that did not, as proven by two Mann-Whitney U-tests at $\alpha = 0.05$ that delivered p-values of .202 and .392, respectively. However, there was a significant difference in AOI hits on the email body. A Mann-Whitney U-test delivered values of $z = 27738.00$, $p < .001$ and $r = .22$, showing a significant difference at $\alpha = 0.05$ with small effect. Figure 8 shows that participants without PAS who correctly identified a phishing email had less AOI hits on the email body than participants without PAS who fell for a phishing email. This indicates that users with the ability to correctly identify a phishing attempt need less time to extract the relevant information from the email body.

But the interesting results happen in the group with PAS. In order to test H4, as before, for all 12 phishing emails it was studied whether AOI hits differ between participants

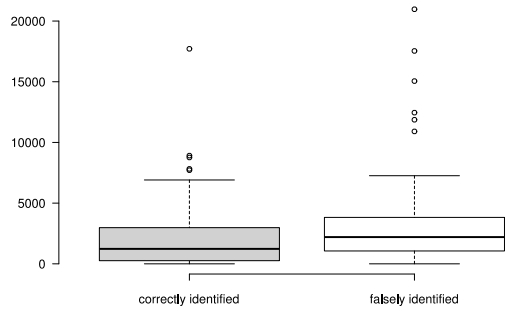

Figure 8. AOI hits on the email body for participants without PAS that correctly identified the phishing email and those that did not.

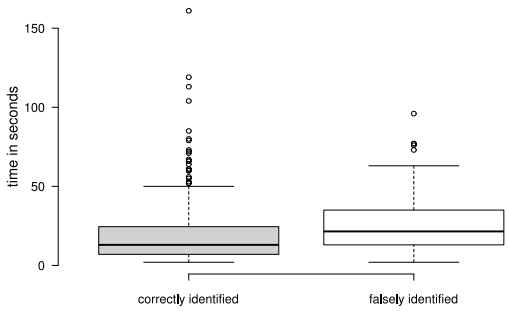

Figure 9. Time needed to identify phishing emails for users with PAS.

who correctly identified the email and participants who did not. Since samples are again non-normally distributed, Mann-Whitney U-tests at $\alpha = 0.05$ were employed, see Table II below.

TABLE II. MANN-WHITNEY U-TEST RESULTS FOR AOI HITS BETWEEN PARTICIPANTS WHO CORRECTLY IDENTIFIED A PHISHING EMAIL VERSUS THOSE THAT DID NOT, IN THE GROUP WITH PAS.

	z	p	r
Subject and Sender	32799.00	$< .001$	.18
Email Body	35047.00	$< .001$	.24
Attachment	25642.00	.973	.00
PAS total	31171.50	$< .001$	.14
PAS contained attachments	30966.50	$< .001$	.16
PAS contained links	31016.00	$< .001$	.14
PAS sender address	30642.50	.001	.13

PAS contained attachments, links and sender address refer to the specific areas in the PAS where the phishing markers are highlighted. They are included separately here to allow for a more detailed evaluation.
These results show significant differences with small effect between the two groups in the number of AOI hits on all AOIs except for the attachment.

TABLE III. MEDIANS OF AOI HITS FOR THE GROUP WITH PAS THAT IDENTIFIED A PHISHING EMAIL CORRECTLY VERSUS THE GROUP WITH PAS THAT FELL FOR THE PHISHING ATTEMPT.

	Median Group Correct	**Median Group False**
Subject and Sender	103.50	457.00
Email Body	838.00	2034.00
PAS total	242.50	439.50
PAS contained attachments	0.00	26.00
PAS contained links	0.00	74.50
PAS sender address	45.00	114.50

It can be seen in Table III that in the group with PAS, participants who correctly identified a phishing email spent less time looking at phishing markers than participants who fell for the phishing attempt. In the group without PAS, this effect could only be seen in regards to the email body. This indicates that, while the PAS does not make all users more effective in identifying phishing emails, it does make users more efficient that have the sufficient knowledge to identify phishing attempts. A likely interpretation of this result is that users who are proficient in identifying phishing emails benefit from the clarity and overview provided by the PAS and are thus enabled to make their decision faster. The r-values for all phishing markers in which a significant difference was found are of similar size, indicating that the PAS highlights all necessary information except for attachments equally. The same effect can be seen when only looking at the processing time per email instead of the individual AOIs, as seen in Figure 9.

Users with the PAS spent less time on phishing emails that were identified correctly as opposed to phishing emails that were identified falsely. This indicates again a gain in efficiency through the PAS when users are already confident in their decision, but no gain in visibility for individual phishing markers. Hypothesis 4 can thus be accepted, but it remains to say that no satisfying answer to RQ5 could be found. While there is a significant difference in time between participants that recognise a phishing attempt and those that do not in the group with PAS, no definitive statement can be made on which phishing markers are overlooked when a user falls for a phishing attempt.

Overall, participants fell for a phishing attempt with a suspicious attachment in 19% of cases, for a phishing attempt containing a suspicious link in 15% of all cases and for a phishing attempt containing an injunction to send money or other items of value in 11% of cases. This highlights the dangers of phishing attacks and the susceptibility of users to fall especially for phishing attempts with attachments. To counteract this effect, an organization-wide attachment blocker can be used, only allowing attachments of certain file types. To prevent users from clicking on a phishing link, a generic phishing warning on emails containing links is effective [3]. Participants were least likely to fall for a phishing attempt involving an injunction to send items of value, however, at 11% the failure rate is still quite high. Here, again, it is crucial to invoke warnings on emails from external senders [3].

Similar to the findings in [3], this study could not find a significant difference in phishing detection between participants who stated they had already taken part in IT-security training and those who did not. There is a need to distinguish here between this study and the referenced paper: one examined voluntary, contextual training, while the other only asked if participants had ever taken part in any IT-security training, however long ago. Still, this results highlights the necessity for further study to achieve innovative, tailored, and effective training methods.

A. Summary of results

The use of an additional phishing awareness system did not improve phishing recognition or the efficiency of phishing recognition. However, using the PAS leads to less time needed to gather information regarding the attachments of a suspicious email. Additionally, users with the PAS who correctly identified a phishing email spent less time looking at all phishing markers except for attachments, compared to users with the PAS who fell for the phishing attempt. This result could not be seen in the group without PAS, indicating that its existence helps users who already have the necessary knowledge to identify phishing emails to make their decision faster. By adapting the tool using existing human-computer interaction guidelines, one can hope to achieve a benefit to all users, not just the experts, in the future. Previous phishing training was proven to have no effect on how likely a participant is to fall for a phishing attempt. Participants fell most often for phishing emails with suspicious attachments and least often for phishing emails with injunctions to send items of value.

V. Usability Results

Although the PAS prototype did not lead to an overall improvement in the effectiveness or efficiency of detecting phishing emails, it did help specific user groups identify phishing markers more quickly. Generally, participants rated the tool's usability as relatively good. The SUS questionnaire yielded an average score of $M = 75,15$, indicating a good usability score. The UEQ-S confirmed these findings, with the measured pragmatic quality — strongly related to usability [21] — scoring $M = 1,53$, indicating a "Good" to "Above Average" result. However, the hedonic quality, which measures non-task-related experience, scored "Below Average". The detailed results of the UEQ-S are provided in Figure 10. This suggests that while the tool meets users' functional requirements, it does not deliver an outstanding experience.

Furthermore, analyzing the SUS scores revealed that users rating the usability as good (68 and higher, as defined by [22]) were able to correctly identify more emails in total and phishing emails compared to users rating the usability of PAS as below average. The p-values and effect sizes indicated statistically significant differences between the two groups with a medium effect. The results of the Mann-Whitney U-test can be seen in Table IV.

Both the UEQ-S and SUS results should be further investigated to explore potential connections between the efficiency and effectiveness of phishing detection and the perceived usability and user experience.

TABLE IV. Mann-Whitney U-test results between users giving a good usability rating and users giving a below average usability rating.

	z	p	r
Total number of correctly identified emails	157.50	.018	.33
Number of correctly identified phishing emails	155.00	.014	.35

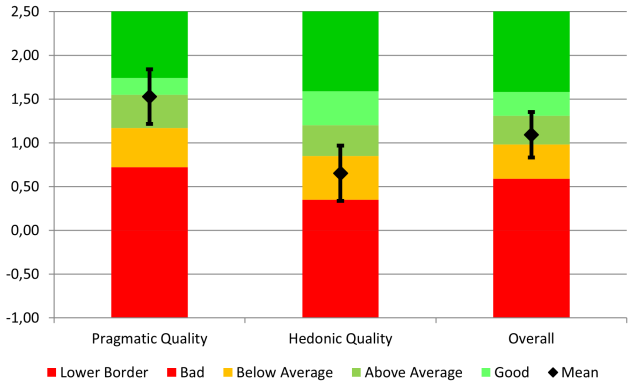

Figure 10. Results of the UEQ-S Questionnaire.

VI. Limitations

The uneven distribution in gender, prior training and exposure to phishing emails between the two groups is to be considered a limiting factor. The effect of prior training is thought to be negligible, since no effect of training could be found in this study. However, only 57% of participants in the PAS group receive phishing mails daily or several times a week, compared to 71% in the group without PAS. This could certainly be an influence as to why no difference in effectiveness between the two groups could be found. Secondly, none of the usability and UX-related questions addressed the participants trust in the tool. Scepticism towards an unfamiliar tool may have been a factor that lead to no measurable difference in efficiency being found between both groups. Understanding and addressing the human elements can enhance the overall effectiveness of security awareness campaigns, ensuring that users are better prepared to recognize and respond to potential threats [23]. Lastly, while the varied recording locations allowed for a diverse and representative set of participants, this also meant that external factors unique to each location could influence the data quality. These include differing levels of natural light, varying background noise levels and differences in posture due to variations in tables and seating heights. As a result, the data quality cannot be compared to eye-tracking studies conducted under laboratory settings. To determine whether these factors influenced the final results, a smaller follow-up study could be conducted to

compare the study design in both controlled and uncontrolled environments.

VII. Conclusion and Future Work

This study demonstrates that relying solely on task-related efficiency and effectiveness metrics, such as the number of correctly identified emails and completion time, does not provide a complete picture of the effectiveness of cybersecurity tools. Significant insights come from understanding users subconscious interactions with the system, which eye-tracking technology can reveal. Understanding these interactions is crucial because systems are only as secure and robust as their weakest link. The collected eye-tracking data is comprehensive and warrants further examination in subsequent studies.

Especially RQ4 and RQ5 have shown that the PAS tool was able to help users with sufficient knowledge in detecting phishing markers more quickly. If an influence on one group can be measured, it is likely that the system can be adjusted to help other user groups as well. The PAS prototype could be tailored to fit a wider audience by, for example, supplying additional information that users susceptible to phishing attacks might need. Given the broad debate on the effectiveness of security training and tools [3][4][23], the fact that the evaluated PAS tool was able to support specific user groups and received positive ratings from users can be considered a success. Further developing the tool to be more user-centered could not only lead to a higher perceived hedonic quality but also an increase in overall effectiveness for all user groups [24].

Cybersecurity and information security depend on robust technological systems, physical defenses against attacks, and the security awareness of end users. While phishing attacks using harmful attachments can be effectively countered with suitable blockers, phishing attacks targeting the end user persist. While phishing training alone seems not to be the sole solution to phishing prevention, the problem of security awareness needs to be addressed in some form. A combination of suitable tools and adequate training on the use of these tools, as well as on the broader topic of security awareness, could help companies reduce the total number of successful phishing attacks.

Acknowledgments

This study was conducted as part of the EU-funded EDIH *Digital Innovation Ostbayern (DInO)*. DInO is funded by the European Union (Project Reference 101083427) and the European Funds for Regional Development (EFRE) (Project Reference 20-3092.10-THD-105). The eye-tracking study was approved by the Joint Ethics Committee of the Bavarian Universities (GEHBa) with the reference number GEHBa-202312-V-155-R.

Data

The eye-tracking and questionnaire data collected and evaluated in this study is free to use and can be found on Zenodo under the following link doi.org/10.5281/zenodo.13171791.

References

[1] N. S. Sulaiman *et al.*, "Cyber–information security compliance and violation behaviour in organisations: A systematic review," *Social Sciences*, vol. 11, no. 9, p. 1, 2022, ISSN: 2076-0760. DOI: 10.3390/socsci11090386.

[2] ISO27001, "Information technology, security techniques, information security management systems, requirements," *International Organization for Standardization ISO*, 2005.

[3] D. Lain, K. Kostiainen, and S. Capkun, "Phishing in organizations: Findings from a large-scale and long-term study," *2022 IEEE Symposium on Security and Privacy (SP)*, p. 9, 2022.

[4] A. Heinemann and G. Schembre, "Zur Wirksamkeit von Security Awareness Maßnahmen [on the effectiveness of security awareness measures]," ger, in *DACH Security Tagungsband 2017: Bestandsaufnahme, Konzepte, Anwendungen, Perspektiven*, P. Schartner and A. Baumann, Eds., Klagenfurt (Österreich): Alpen-Adria-Universität, 2017, ISBN: 978-3-00-057290-6.

[5] J. L. Orquin and K. Holmqvist, "Threats to the validity of eye-movement research in psychology," *Behavior Research Methods*, vol. 50, no. 4, pp. 1645–1656, Aug. 2018, ISSN: 1554-3528. DOI: 10.3758/s13428-017-0998-z.

[6] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," vol. 2, Jan. 2007.

[7] L. Grabinger, F. Hauser, C. Wolff, and J. Mottok, "On eye tracking in software engineering," *SN Computer Science*, vol. 5, no. 6, p. 729, Jul. 26, 2024, ISSN: 2661-8907. DOI: 10.1007/s42979-024-03045-3.

[8] L. Huang, S. Jia, E. Balcetis, and Q. Zhu, "Advert: An adaptive and data-driven attention enhancement mechanism for phishing prevention," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 2585–2597, 2022. DOI: 10.1109/TIFS.2022.3189530.

[9] J. Kävrestad *et al.*, "Evaluation of contextual and game-based training for phishing detection," *Future Internet*, vol. 14, no. 4, 2022, ISSN: 1999-5903. DOI: 10.3390/fi14040104.

[10] J. McAlaney and P. J. Hills, "Understanding phishing email processing and perceived trustworthiness through eye tracking," *Frontiers in Psychology*, vol. 11:1756, 2020, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2020.01756.

[11] F. Pietrantonio *et al.*, "Investigating gaze behavior in phishing email identification," in *2023 7th Network Traffic Measurement and Analysis Conference (TMA)*, 2023, pp. 1–4.

[12] M. Clark, I. Ruthven, and P. O. Holt, "Perceiving and using genre by form – an eye-tracking study," *Libri*, vol. 60, no. 3, pp. 268–280, 2010. DOI: doi:10.1515/libr.2010.023.

[13] M. Clark, I. Ruthven, P. O. Holt, D. Song, and S. Watt, "You have e-mail, what happens next? tracking the eyes for genre," *Information Processing & Management*, vol. 50, no. 1, pp. 175–198, 2014, ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2013.08.005.

[14] R. Yang, K. Zheng, B. Wu, C. Wu, and X. Wang, "Prediction of phishing susceptibility based on a combination of static and dynamic features," *Mathematical Problems in Engineering*, p. 2 884 769, 2022. DOI: https://doi.org/10.1155/2022/2884769.

[15] L. Jaeger and A. Eckhardt, "Eyes wide open: The role of situational information security awareness for security-related behaviour," *Information Systems Journal*, vol. 31, no. 3, pp. 429–472, 2021. DOI: https://doi.org/10.1111/isj.12317. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/isj.12317.

[16] K. Pfeffel, P. Ulsamer, and N. H. Müller, "Where the user does look when reading phishing mails – an eye-tracking study," in *Learning and Collaboration Technologies. Designing Learning*

Experiences, P. Zaphiris and A. Ioannou, Eds., Cham: Springer International Publishing, 2019, pp. 277–287, ISBN: 978-3-030-21814-0.

[17] N. Xu, J. Fan, and Z. Wen, "Email reading behavior-informed machine learning model to predict phishing susceptibility," *Lecture Notes in Computer Science, vol 14509. Springer, Singapore*, pp. 579–592, 2024.

[18] Y. Abdrabou *et al.*, *Revealing the hidden effects of phishing emails: An analysis of eye and mouse movements in email sorting tasks*, 2023. arXiv: 2305.17044.

[19] C. Blake, "Eye-Tracking: Grundlagen und Anwendungsfelder [Eye-Tracking: Foundations and Applications]," ger, in *Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft*, W. Möhring and D. Schlütz, Eds., Wiesbaden: Springer Fachmedien Wiesbaden, 2013, pp. 367–387, ISBN: 978-3-531-18776-1. DOI: 10.1007/978-3-531-18776-1_20.

[20] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)†," *Biometrika*, vol. 52, no. 3-4, pp. 591–611, Dec. 1965, ISSN: 0006-3444. DOI: 10.1093/biomet/52.3-4.591. eprint: https://academic.oup.com/biomet/article-pdf/52/3-4/591/962907/52-3-4-591.pdf.

[21] M. Hassenzahl, A. Platz, M. Burmester, and K. Lehner, "Hedonic and ergonomic quality aspects determine a software's appeal," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '00, The Hague, The Netherlands: Association for Computing Machinery, 2000, pp. 201–208, ISBN: 1581132166. DOI: 10.1145/332040.332432.

[22] J. R. Lewis and J. Sauro, "Item Benchmarks for the System," en, *Journal of Usability Studies*, vol. 13, no. 3, pp. 158–167, 2018.

[23] M. Bada, A. M. Sasse, and J. R. Nurse, "Cyber security awareness campaigns: Why do they fail to change behaviour?" *arXiv preprint arXiv:1901.02672*, 2019.

[24] M. A. Sasse, S. Brostoff, and D. Weirich, "Transforming the 'weakest link' — a human/computer interaction approach to usable and effective security," *BT Technology Journal*, vol. 19, no. 3, pp. 122–131, Jul. 1, 2001, ISSN: 1573-1995. DOI: 10.1023/A:1011902718709.