Towards an Empirical Study to Determine the Effectiveness of Support Systems against E-Mail Phishing Attacks

Katharina Schiller
katharina.schiller@hof-university.de
Hof University of Applied Sciences,
Institute of Information Systems
Hof, Bavaria, Germany

Florian Adamsky
florian.adamsky@hof-university.de
Hof University of Applied Sciences,
Institute of Information Systems
Hof, Bavaria, Germany

Zinaida Benenson
zinaida.benenson@fau.de
Friedrich-Alexander-Universität
Erlangen-Nürnberg (FAU)
Erlangen, Bavaria, Germany

ABSTRACT

E-mail phishing attacks are still the number one gateway for attackers. Even when the patch level of a network is up to date, if one employee clicks on a link in a phishing e-mail and enters their credentials on a malicious website or downloads malware, the whole organization might get compromised. Anti-phishing support systems highlight different aspects of an e-mail to help users to detect phishing e-mails. However, little is known about their effectiveness, especially in comparison to each other. This paper presents our experimental design to investigate the efficacy of various support systems. For this purpose, we created a fictional scenario and an interactive tool to display e-mails. In addition, we present our preliminary study with the first results to classify test e-mails in different difficulty levels that serve as a basis for our main study.

ACM Reference Format:
Katharina Schiller, Florian Adamsky, and Zinaida Benenson. 2023. Towards an Empirical Study to Determine the Effectiveness of Support Systems against E-Mail Phishing Attacks. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3544549.3585658

1 INTRODUCTION

Hardly a day goes by that we do not receive e-mails that try to lure us into clicking on links, entering our credentials on websites, or installing malicious software. This social engineering attack is known as *phishing* and is one of the biggest security threats. The Anti-Phishing Working Group (AWPG) [4] states in their Q2 report of 2022: "[...] *this is the worst quarter for phishing that APWG has ever observed*". According to the ESET report [14], in 2021, the number of e-mail phishing attacks increased by 7.3 %. Additionally, a CISCO report from 2021 [8], shows that phishing accounts for approximately 90 % of data breaches. Phishing attacks are not restricted to e-mails, but since e-mail is still the most used communication channel in the business area, it comprises 96 % of all phishing attacks [29]. Additionally, e-mail is by default not authenticated, and attackers can easily spoof a sender of an e-mail.

As major phishing attacks have been widely publicised, users should be aware of e-mails as an attack vector but e-mails remain a relevant threat [19]. Bada et al. [5] note that users can understand the danger, but are not necessarily motivated to change their behaviour. Further, attackers keep using new techniques [28] that make it very difficult to recognise threats, and even trained users fall for them.

Although there are plenty of technical countermeasures against e-mail phishing, such as DomainKeys Identified Mail (DKIM), Sender Policy Framework (SPF), and Domain-based Message Authentication, Reporting and Conformance (DMARC), these are often hard to implement, and break existing workflows such as mailing lists and have a high maintenance, such as changing cryptographic keys periodically [16, 23]. In addition, probabilistic technical countermeasures such as filters have to balance their false positive and false negative rates, such that they cannot block all phishing e-mails, especially when it comes to spear-phishing.

Support systems are based on technical countermeasures but do not block suspicious e-mails. Instead, they provide hints and highlight e-mail elements that could indicate phishing, trying to support users in their decisions. For example, they may show if an e-mail comes from an internal or external Mail Transfer Agent (MTA). To the best of our knowledge, no study has yet compared the effectiveness of different support systems.

The contributions of this paper are as follows: We

- present our current study design to compare the effectiveness of various support systems;
- present the results of a pre-study to determine the difficulty of the test e-mails and discuss lessons learned.

2 BACKGROUND AND RELATED WORK

Kumaraguru et al. [21] divide anti-phishing measures into three categories. These are referred to as *Technical Countermeasures, User Education and Training*, and *Support Systems* and are described in the following in more detail.

2.1 Technical Countermeasures

Technical countermeasures are usually hidden from the user and directly performed on the server side. For example, SPF [18] verifies if an e-mail is sent from the correct MTA. The mail servers' IP addresses allowed to send e-mails are stored in a TXT record of that domain. However, SPF only checks the envelope-from address, *not* the address from the `FROM` field. DomainKeys Identified Mail [9], on the other hand, verifies the senders' domain and integrity using asymmetric cryptography. The public key of the MTA is also stored in a TXT record of the domain. The MTA uses its private key to create a signature of the body of the e-mail, including the `FROM` field and attaches it to the e-mail header. The receiver

CHI EA '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3585658

MTA can then check the signature with the published public key. However, nothing prevents the adversary from not using DKIM. This is where DMARC [9] comes into play. DMARC uses SPF and DKIM and checks that the FROM field is identical to the envelope-from field, which would prevent e-mail spoofing. Unfortunately, this breaks mailing lists that change header fields such as TO, FROM and SUBJECT. For this reason, the adoption rate of this technology is still low [16].

2.2 User Education and Training

Many companies and organizations are aware of the threat caused by phishing e-mails and try to counter it by training their members and increasing their awareness. A classical education way is in-class training courses held by an instructor. Stockhardt et al. [27] show that this approach is effective but time-consuming for participants. On the other hand, text-based training [27] imparts knowledge, for instance, via websites, newsletters or printed posters and leaflets. The advantage is that users can work through them individually at their own pace and according to their prior knowledge. Furthermore, there are gamification approaches with interactive applications [25, 32, 33]. Related to the previous two methods is embedded training. In this case, users receive a fake phishing e-mail that displays a learning website that can contain text-based training or gamification elements when clicked. Often the click-through rate is measured. That shows how many recipients clicked on the link. While studies [7, 17, 20] show the effectiveness of such phishing simulations, they mainly focus on a relatively short period. With longer observation, the click-through rate rises again [6]. Furthermore, phishing simulations also have disadvantages [31]: Firstly, technical measures have to be adapted so that fake phishing e-mails do not get blocked, which opens new attack vectors. Secondly, organizations "trick" their members by phishing simulations, which could negatively impact the trust relationship. Finally, members might fear consequences due to mistakes, leading to an increase in false positives.

2.3 Support Systems

A middle ground between the two presented approaches is the so-called support system. On the one hand, they are based on technical solutions; on the other hand, they rely on users' decision-making. They do not block suspicious e-mails or websites directly but instead mark them or highlight any irregularities that are found. This also prevents false positives from being hidden from the user [3]. We distinguish two kinds of support systems: browser-related and Mail User Agent (MUA)-related.

2.3.1 Browser-related Support Systems. These support systems are only visible to the user after they click a link in an e-mail. For example, *indicator icons* may show a lock symbol next to the URL in the browser window, which indicates whether the established server connection is secure. Various studies [10, 11, 15, 34] found that users do not perceive or do not understand the icons and instead use the content on the website to indicate its authenticity. Another approach addresses the problem of complex URLs that are difficult for users to understand. Albakry et al. [2] show that users assumed a URL belonged to a certain organization if its name was included somewhere in that URL. At the same time, they ignored the position and other relevant aspects that indicate whether the URL actually belongs to that organization. *Domain highlighting* highlights visually the relevant parts of a URL. Studies [13, 22] show that this technology is not sufficient since users mainly focus on the content area and overlook the highlighted domain. Therefore, warnings that interrupt the flow of the user and require active intervention are helpful [12]. If the technical countermeasures suspect a malicious website, the user is warned by a full-screen message from the browser. However, the user can continue to the website or leave. These are *active warnings* as users are interrupted in their task flow and are forced to take action to continue. Usually, the warning offers the user two options *continue* or *cancel*, which indicates a recommendation through different visual design and wording [1, 12]. In contrast, browsers also include *passive warnings* that are not screen-filling, and the actual website is already displayed in the background. Egelman et al. [12] compare passive and active warnings. Their results show that passive warnings are ineffective and make no difference to no warning message.

2.3.2 MUA-related support systems. Another way to support users is to warn them *before* they click a link in an e-mail. MUA-related support systems provide hints that can be placed above an e-mail or directly in the content area. The problem with HTML e-mails is that the writer can use anchors; therefore, the URL is hidden behind the anchor text, such as "Click here". Usually, desktop MUAs show a *tooltip* next to the mouse cursor or at the bottom of the window if a user hovers over the link. These tooltips show the actual URL the link leads to. However, the reader of an e-mail can easily oversee the URL or does not know the structure of an URL; therefore, the reader cannot assess if it is a legitimate URL. In [30], Volkamer et al. examine a novel tooltip system which they call TOoltip-poweRed Phishing E-Mail DetectiOn (TORPEDO). Comparable to domain highlighting, the domain is visually highlighted in the tooltip. In order to draw attention to the tooltip, clicking on the link is delayed by a few seconds. Their study results are promising since TORPEDO helps users to identify phishing e-mails.

Another method that has not yet been investigated in the literature, but is widely used in practice, is the marking of e-mails with *external markers*. Depending on the MTAs of the receiver, the term "external" in front of the name of the sender or the e-mail is otherwise highlighted as external. Especially if an e-mail from within the organization is marked as external, it is a warning signal.

Similarly to browser warnings, MUA can also display warning messages to the user. Petelka et al. [24] compare different warning methods and show that placing the warning directly next to the link is more effective than as a warning banner on top of the e-mail. Labels or sidebars are also a form of support system. These are available as add-ons and must be installed first. For some of them, the user must click on *check e-mail* to get feedback from the support system. This feedback gives users transparent information about the e-mail and tags it to indicate its trustworthiness. To the best of our knowledge, no studies compare the effectiveness of different types of MUA-related support systems.

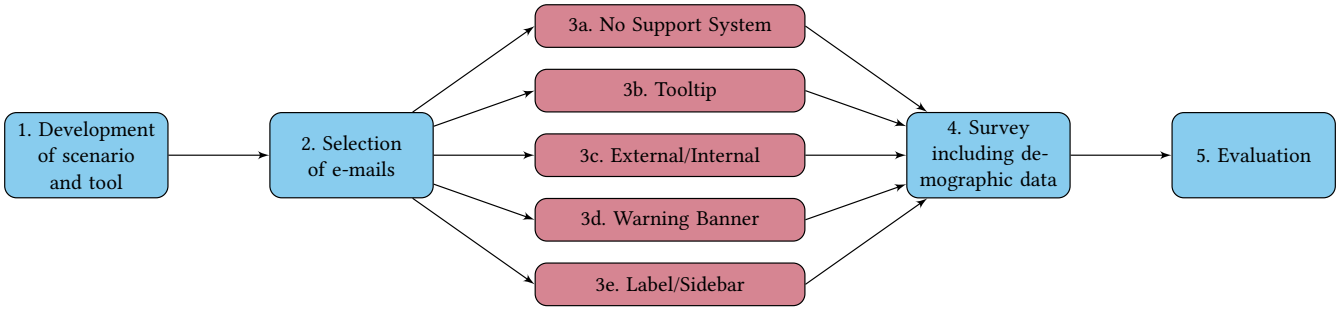


Figure 1: The approach of our study design. The blue boxes are work steps and the red boxes are the different support systems that we will evaluate with the same e-mails.

3 STUDY DESIGN

Our objective is to evaluate the effectiveness of different support systems. For this purpose, we need various e-mails where we can integrate the support systems. These include phishing e-mails but also legitimate e-mails. While other studies often use e-mails without testing their difficulty, we evaluate them in advance to determine which impact the support system has on the difficulty. This means we examine how easy or difficult it is to recognise the e-mails as phishing or legitimate, and based on this, make a selection of various difficult e-mails. Figure 1 shows our approach to the study design. The selection of e-mails is the first step in our study design. We show our exact approach and first results in Section 3.2.

Our study focuses on MUA-related support systems; browser-related ones are out of scope since there are already studies [1, 12, 15] focusing on that. The following categories cover most of the MUA-related support systems we have found through our literature research and looking at actual implementations: Tooltip [30], External Marker, Warning Banner, and Label/Sidebar are referenced in Figure 1 as 3b–3d, respectively. As mentioned in Section 2.3, support systems can highlight e-mails in different ways. Support systems also mark legitimate e-mails as being valid or highlight them in green. Alternatively, they present different degrees of danger. However, not all support systems show these degrees, such as the External Marker (3c), where is only one variant. For this reason, we are using the medium variant for all support systems that provide different degrees. We also considered false positives. In reality, false positives where legitimate e-mails are highlighted as phishing might occur. However, for our study, we do not consider false positives not to influence the participants' trust in a support system during the study. That means that all phishing e-mails are highlighted with the medium variant of the support system, and all legitimate e-mails are not. Exceptions are the External Marker, where all external e-mails are highlighted, and the Tooltip (3b), where only e-mails with a link can be highlighted.

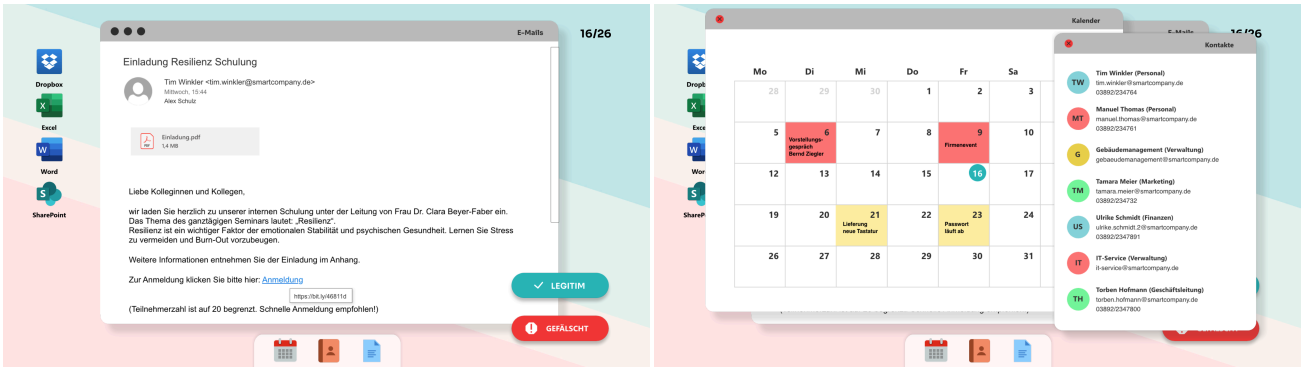
We want to conduct a between-subjects online study. This means we divide the participants into different experimental groups, each seeing one support system and a control group without a support system. Participants must then classify a mixture of phishing e-mails and legitimate e-mails as phishing or legitimate. A field study would require that the participant groups have simultaneously activated different support systems and that we measure the effectiveness using the click-through rate. Phishing simulations in companies use this principle to increase awareness.

In an online study, we can collect further information from the participants; see work step 4 in Figure 1. Demographic information can help us to classify the results better. In a final questionnaire, we would like to go into more detail about the particular support system. Among other things, we want to ask whether the participants noticed the support system and whether they found it helpful. In addition, we want to gain insights into whether the participants understand the support system and the information communicated therein and how these can be improved.

3.1 Scenario and Tool

In a company, an employee knows their colleagues, upcoming deadlines or software in use. So we have to set this context with other relevant information in our study. Depending on the context of the e-mail, participants can, for example, rate it by whether they know the sender or whether it mentions a familiar appointment. For our study, we created an employee of the Human Resources (HR) department of an IT company called *Smartcompany* as a fictional character named *Alex*. The participants are then asked to slip into the role of Alex and classify the e-mails.

For this purpose, we developed an interactive tool that displays all relevant information on a simplified desktop interface seen in Figure 2. Our interactive tool shows e-mails in random order. An e-mail counter is shown in the top right corner of the screen. On the opposite side of the screen, four application icons are visible that the company uses in our fictional scenario and are relevant to the classification of e-mails. The actual e-mail is located in the middle, in a highly simplified MUA, based on Outlook. We decided to use Outlook as a basis, as this is a widely used MUA. The functionality of the tool is limited to allowing participants to scroll to view long e-mail texts and hover over links or buttons to view the actual URL. There is no possibility of replying to the e-mail or writing e-mails oneself. There are two buttons available for the classification of the e-mails: a blue highlighted button with a tick icon labelled *legitimate*, if the participant thinks an e-mail is trustworthy; a red highlighted button with an exclamation mark icon labelled *fraudulent*, if a participant thinks the e-mail is phishing. As we mentioned earlier, we created a fictional context. Unfortunately, it was too much



(a) Tool (b) Calendar and Contacts

Figure 2: Subfigure (a) shows a screenshot of the tool we developed to determine the difficulty of legitimate and phishing e-mails. It shows a legitimate e-mail. Subfigure (b) depicts the user's calendar and e-mail address book, which is part of the tool to set the context. It shows information about our fictional scenario that helps to classify the e-mails correctly.

information to memorise to give to participants in a listing at the beginning. For that reason, we included this information in our interactive tool. At the bottom of the screen, participants can find three icons for *calendar*, *contacts*, and *job offers*. Job offers refer to the fact that our character works in the HR department and shows the job title of an, in our fictional context, actually advertised job. The calendar and contacts can be seen in Figure 2. The calendar shows four entries: two appointments, highlighted red, and two relevant dates, highlighted yellow. Two entries are in the past, and two are in the future. The 16^{th} is highlighted as the current day. Last week, the character had an interview with an applicant and a company event on Friday. In the coming week, the character expects a delivery with new hardware and the current password for the account expires. The last icon is the contacts icon which opens a simplified address book. It shows a listing of various *Smartcompany* employees and their departments. Furthermore, their e-mail addresses and telephone numbers are listed. There are also two function e-mail addresses in the address book for facility management and IT-Service.

3.2 E-Mail Selection

For our study, we need e-mails with different difficulty levels. In this way, we want to prevent all e-mails from being correctly classified without any problems and without a support system. At the same time, it must still be possible to classify the e-mails correctly. Another point we consider is the ratio between fraudulent and legitimate e-mails. For example, the following studies [11, 22, 30] used an even or approximately even ratio between fraudulent and legitimate e-mails. Whereas other studies [21, 24][1] used a higher number of legitimate e-mails. We were of the opinion that a higher amount of legitimate e-mails represents a real-world scenario, meaning we want to use three phishing and twelve legitimate e-mails. Another phishing e-mail is a tutorial to let the participant familiarise themselves with the system. We used the *phish scale* by Steves et al. [26] and the e-mails they used in their study as a basis for our phishing e-mails. They present a scale to rate phishing e-mails regarding their difficulty level.

The phish scale consists of two main components: the cues for phishing and the premise alignment. Cues are hints in an e-mail that may indicate to be fraudulent. Premise alignment describes how it suits the context of the recipient. Both scores for cues and premise alignment are combined and lead to the overall rating of e-mails. Initially, we used the phish scale to create both phishing and legitimate e-mails. For the legitimate e-mails, we tried to use the phish scale inverted, which means, all legitimate e-mails should be difficult phishing e-mails on the scale. After we created the e-mails, two persons rated them independently on the phish scale. Afterwards, they discussed and adjusted the ratings. One of the problems both persons mentioned was that the evaluation of premise alignment was somewhat subjective. However, the majority of e-mails had the same rating. In our opinion, the number of necessary cues according to the phish scale is too high. Depending on the length of the e-mail text, it is challenging to include all necessary cues. We think the phish scale cues need weighting to better reflect the actual difficulty of an e-mail. Because of this, we added more e-mails that we did not rate on the phish scale.

Generally, we tried to find examples that could be received in a similar form as phishing e-mails and fit our scenario. We also tried to adapt the legitimate e-mails to the scenario. Thus, many e-mails were sent by colleagues or departments of the company. But since we also wanted to test the External Marker, e-mails from outside the company had to be included. Altogether, we created 26 e-mails as a selection for the main study—11 phishing and 15 legitimate ones. A broad summary of all e-mails can be seen in Appendix. All phishing e-mails can be found in Appendix as well. Since we conducted the study in Germany, the e-mails are in German language. Besides, we made another legitimate e-mail, which serves as a tutorial for the participants in our pre-study, as will be described in the next section.

[1] Kumaraguru et al. [21] actually used an even ratio for their gamification approach, but a higher number of legitimate e-mails in other parts of their study.

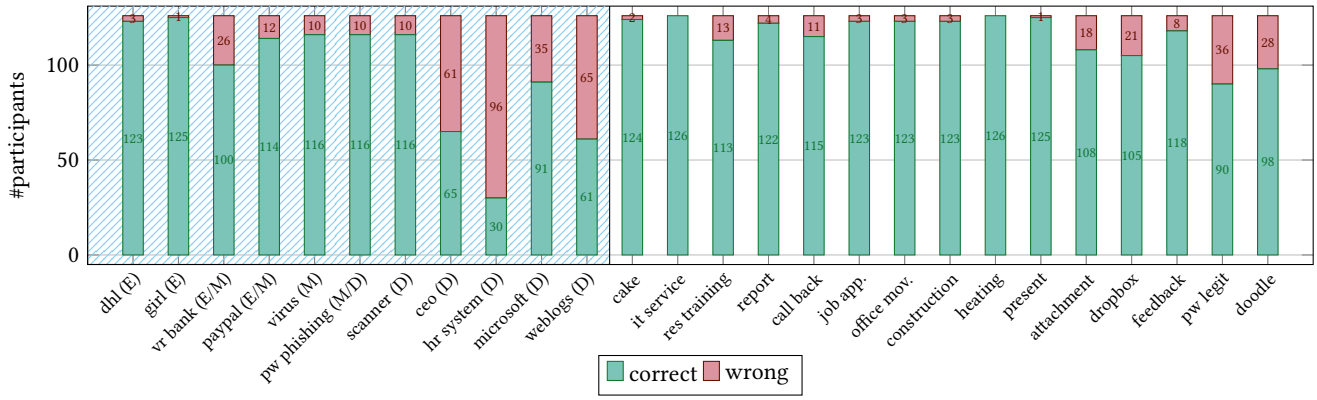


Figure 3: Results of our study with 126 participants. The bars on the left side with a hatched background are phishing e-mails, the others on the right side are legitimate ones. The lower green bars represent the correct, the upper red bars represent the incorrect classifications. The character in the brackets at the label's end represents our assessment. For the phishing e-mails, the *(E)* indicates an easy e-mail, an *(M)* a medium one, and a *(D)* a difficult one.

4 EVALUATION OF E-MAIL DIFFICULTY

We conducted an online-pre-study to select the appropriate e-mails for our study. As far as we know, other studies do not describe if they tested their e-mails in advance. We wanted to ensure that the e-mails could be identified correctly and were from different difficulty levels. For this purpose, we used the German crowdsourcing marketplace clickworker.de[2].

4.1 Pre-Study Design

The participants received a link to our tool. A welcome page provided them with relevant information about the study and indicated an approximate duration of 15 min. An additional legitimate e-mail serves as a tutorial at the beginning and is not included in the evaluation. When participants click start, they first see an interactive tutorial with the tutorial e-mail described earlier. The participants can repeat the steps in the tutorial and classify the e-mail several times in different ways. This is not possible with the other 26 e-mails. It also shows whether the classification on the tutorial e-mail was correct. The tutorial uses animations to show the participants how to use the tool and where to find the relevant information. An animation is used to hover over a link, but this is not explicitly pointed out in order to influence participants who would not do so in real life. To ensure that the tool is usable, we tested it in advance with various people, including the elderly. After the tutorial is complete, participants see all other 26 e-mails in random order and have to classify them using the two buttons. In this pre-study on e-mail difficulty, the participants only received general feedback on how many of the e-mails they classified correctly. In the end, participants were forwarded a questionnaire about their demographic data. All participants received remuneration after completing the study. We tracked the overall time they needed for the study and how long they took for each e-mail. We also saved when someone clicked on the *calendar*, *contacts*, *job offers* or hovered over a link. However, hovering over a link can also happen by accident while reading the e-mail.

[2]https://www.clickworker.de/

4.2 Pre-Study Results

Figure 3 shows the results of our study with 126 participants. We excluded four of 130 results from the evaluation because they ended the survey in a few seconds or always clicked the same button. For this reason, we consider 126 participants. Eighty-three participants were male, and 43 participants were female. All participants stated being born between 1946 and 1999. Two participants did not provide information about their technology affinity. Thirty-six estimated themselves as very technology affine, 75 are average, and 13 are barely technology affine. Twenty-five of our participants have prior knowledge in the IT area or are currently active in the field. Ninety-seven negated the question, and four did not answer.

Our results show that most e-mails were correctly classified by over 90 %. Nevertheless, many participants rated the e-mails wrong, which we previously considered as difficult. For example, in the case of the *CEO* phishing e-mail, this is almost half of all participants with 48.4 % and in the case of the *HR System* and *Weblogs* phishing e-mails, more than half with 76.2 % and 51.6 %. Exceptions are the *Scanner* and *Password Phishing* e-mails, where almost all participants were correct. Almost all phishing e-mails that we estimated to be in the medium difficulty range were also correctly classified by over 90 %. One exception is the *VR Bank* e-mail. Here, 20.6 % were wrong. The participants needed an average of 13.65 s per e-mail. They were the fastest with the *Girl* e-mail, with an average of 6.4 s. The legitimate *Attachment* e-mail took the longest, with an average of 21.25 s. This is followed by *HR System*, *Weblogs* and *Microsoft*, each with an average of around 20 s. A total of 77 participants hovered over at least one link. Five participants hovered over every link. We estimated the *HR System* e-mail to be very difficult in advance because the only real indication of phishing is the URL displayed when hovering over the link. This link was hovered by 32 participants (25.4 %). Of these, 37.5 % correctly identified the e-mail as phishing. The rate for participants who did not hover over the link and correctly classified the e-mail is 19.2 %. Sixty participants clicked the contacts icon, 53 the calendar icon and 56 the job offer icon for at least one e-mail.

Our pre-study is an essential part of the study design to test phishing e-mails for their difficulty level and then select them for the actual study. According to our estimation, a selection could have led to the fact that the phishing e-mails are too easy to recognise and the support systems are unnecessary.

5 CONCLUSION AND FUTURE WORK

This study presented our study design to determine the effectiveness of various support systems against e-mail phishing attacks. Additionally, we showed our preliminary results for the selection of the e-mails that we will use for our study. Finally, we described the selection of our phishing e-mails in detail. In the process, we found that the phish scale in its current form without weighting is unsuitable for creating phishing e-mails.

In future work, we will integrate the support systems in the e-mails and engage participants to see if support systems make any difference in the detection rate.

ACKNOWLEDGMENTS

We thank Andrea Schankin for the fruitful discussions that have highly influenced our study design.

REFERENCES

[1] Devdatta Akhawe and Adrienne Porter Felt. 2013. Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness. In *Proceedings of the 22nd USENIX Security Symposium*. USENIX Association, 182–191.

[2] S. Albakry, Kami Vaniea, and M. Wolters. 2020-04-23. What is this URL's Destination? Empirical Evaluation of Users' URL Reading. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, United States, 1–12. https://doi.org/10.1145/3313831.3376168

[3] K. Althobaiti, Ghaidaa Rummani, and Kami Vaniea. 2019. A Review of Human- and Computer-Facing URL Phishing Features. In *Proceedings of the IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 182–191. https://doi.org/10.1109/EuroSPW.2019.00027

[4] Anti Phishing Working Group (APWG). 2022. *Phishing Activity Trends Report*. Technical Report. Anti Phishing Working Group (APWG). https://docs.apwg.org/reports/apwg_trends_report_q2_2022.pdf

[5] Maria Bada, Angela M. Sasse, and Jason R. C. Nurse. 2019. Cyber Security Awareness Campaigns: Why Do They Fail to Change Behaviour? *CoRR* abs/1901.02672 (2019). arXiv:1901.02672 http://arxiv.org/abs/1901.02672

[6] D. Caputo, S. Pfleeger, J. Freeman, and M. Johnson. 2013-08-23. Going Spear Phishing: Exploring Embedded Training and Awareness. *IEEE Security & Privacy* 12 (2013-08-23), 28–38. https://doi.org/10.1109/MSP.2013.106

[7] A. Carella, Murat Kotsoev, and T. Truta. 2017. Impact of Security Awareness Training on Phishing Click-Through Rates. In *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 4458–4466. https://doi.org/10.1109/BigData.2017.8258485

[8] CISCO. 2021. *Cyber Security Threat Trends*. Technical Report. CISCO. https://learn-cloudsecurity.cisco.com/umbrella-library/2021-cyber-security-threat-trends-phishing-crypto-top-the-list#page=1

[9] D. Crocker, T. Hansen, and M. Kucherawy. 2011. *RFC 6376: DomainKeys Identified Mail (DKIM) Signatures*. STD 76. RFC Editor. http://www.rfc-editor.org/rfc/rfc6376.txt

[10] Rachna Dhamija, J. Tygar, and Marti A. Hearst. 2006. Why Phishing Works. In *Proceedings of the 2006 CHI Conference on Human Factors in Computing Systems*. 581–590. https://doi.org/10.1145/1124772.1124861

[11] Julie S. Downs, Mandy B. Holbrook, and Lorrie Faith Cranor. 2006. Decision strategies and susceptibility to phishing. In *Symposium On Usable Privacy and Security*. Association for Computing Machinery, New York, NY, United States, 79–90. https://doi.org/10.1145/1143120

[12] Serge Egelman, L. Cranor, and J. Hong. 2008. You've been Warned: An Empirical Study of the Effectiveness of Web Browser Phishing Warnings. In *Proceedings of the 2008 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/1357054.1357219

[13] Vivian Ives Philipp Erbenich, Daniel Träder, A. Heinemann, and Meltem Nural. 2019. Phishing Attack Recognition by End-Users: Concepts for URL Visualization and Implementation. In *Proceedings of the 13th International Symposium on Human Aspects of Information Security & Assurance (HAISA)*. University of Plymouth, 179–188.

[14] ESET. 2021. *Threat Report T2 2021*. Technical Report. 1–47 pages. https://www.welivesecurity.com/wp-content/uploads/2021/09/eset_threat_report_t22021.pdf

[15] Adrienne Porter Felt, R. Reeder, A. Ainslie, H. Harris, M. Walker, C. Thompson, M. Acer, E. Morant, and Sunny Consolvo. 2016. Rethinking Connection Security Indicators. In *Proceedings of the 12th Symposium On Usable Privacy and Security (SOUPS)*. 1–13.

[16] Hang Hu, Peng Peng, and Gang Wang. 2018-09. Towards Understanding the Adoption of Anti-Spoofing Protocols in Email Systems. In *2018 IEEE Cybersecurity Development (SecDev)* (Cambridge, MA). IEEE, 94–101. https://doi.org/10.1109/SecDev.2018.00020

[17] K. Jansson and R. V. Solms. 2013. Phishing for Phishing Awareness. *Behaviour & Information Technology* 32, 6 (2013), 584–593. https://doi.org/10.1080/0144929X.2011.632650

[18] S. Kitterman. 2014. *RFC 7208: Sender Policy Framework (SPF) for Authorizing Use of Domains in Email, Version 1*. RFC 7208. RFC Editor. http://www.rfc-editor.org/rfc/rfc7208.txt

[19] Katharina Krombholz, Heidelinde Hobel, M. Huber, and E. Weippl. 2015. Advanced Social Engineering Attacks. *Journal of Information Security and Applications* 22 (2015), 113–122. https://doi.org/10.1016/j.jisa.2014.09.005

[20] P. Kumaraguru, Y. Rhee, Steve Sheng, S. Hasan, A. Acquisti, L. Cranor, and J. Hong. 2007. Getting Users to Pay Attention to Anti-Phishing Education: Evaluation of Retention and Transfer. In *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit (eCrime '07)*. https://doi.org/10.1145/1299015.1299022

[21] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason I. Hong. 2010-05. Teaching Johnny not to fall for phish. *ACM Transactions on Internet Technology* 10 (2010-05), 1–31.

[22] Eric Lin, Saul Greenberg, Eileah Trotter, David Ma, and John Aycock. 2011. Does domain highlighting help people identify phishing sites?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/1978942.1979244

[23] Sourena Maroofi, Maciej Korczyński, Arnold Hölzel, and Andrzej Duda. 2021. Adoption of email anti-spoofing schemes: a large scale analysis. *IEEE Transactions on Network and Service Management* 18, 3 (2021), 3184–3196.

[24] 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*. https://doi.org/10.1145/3290605.3300748

[25] Steve Sheng, Bryant Magnien, P. Kumaraguru, A. Acquisti, L. Cranor, J. Hong, and Elizabeth Ferrall-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 (SOUPS)*. https://doi.org/10.1145/1280680.1280692

[26] Michelle Potts Steves, Kristen K. Greene, and Mary Frances Theofanos. 2020-09. Categorizing human phishing difficulty: a Phish Scale. *Journal of Cybersecurity* 6, 1 (2020-09). https://doi.org/10.1093/cybsec/tyaa009

[27] Simon Stockhardt, Benjamin Reinheimer, M. Volkamer, P. Mayer, Alexandra Kunz, P. Rack, and D. Lehmann. 2016. Teaching Phishing-Security: Which Way is Best?. In *Proceedings of the 31st International Conference on ICT Systems Security and Privacy Protection (IFIP SEC 2016)*.

[28] Bill Toulas. 2022-09-13. *Hackers Now Use 'Sock Puppets' for More Realistic Phishing Attacks*. BleepingComputer. https://www.bleepingcomputer.com/news/security/hackers-now-use-sock-puppets-for-more-realistic-phishing-attacks/

[29] Verizon. 2020. *Data Breach Investigations Report*. Technical Report. Verizon. 67 pages. https://www.verizon.com/business/en-gb/resources/reports/dbir/

[30] Melanie Volkamer, Karen Renaud, Benjamin Reinheimer, and Kunz Alexandra. 2017. User experiences of TORPEDO: TOoltip-poweRed Phishing Email DetectiOn. *Computers & Security* 71 (2017), 100–113. https://doi.org/10.1016/j.cose.2017.02.004

[31] Melanie Volkamer, Martina Angela Sasse, and Franziska Boehm. 2020. Analysing simulated phishing campaigns for staff. In *European Symposium on Research in Computer Security*. Springer, 312–328.

[32] Patrickson Weanquoi, J. Johnson, and Jinghua Zhang. 2018-12. Using a Game to Improve Phishing Awareness. *Journal of Cybersecurity Education, Research and Practice* 2018, 2 (2018-12). https://digitalcommons.kennesaw.edu/jcerp/vol2018/iss2/2

[33] Zikai Alex Wen, Z. Lin, Rowena Chen, and E. Andersen. 2019-05. What.Hack: Engaging Anti-Phishing Training Through a Role-playing Phishing Simulation Game. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12. https://doi.org/10.1145/3290605.3300338

[34] Tara Whalen and Kori Inkpen Quinn. 2005. Gathering Evidence: Use of Visual Security Cues in Web Browsers. In *Proceedings of Graphics Interface*. 137–144. https://doi.org/10.5555/1089508.1089532

APPENDIX

Table 1: A description of all phishing e-mails we created for our survey.

Phishing E-mail	Description	Estimated Rating
Weblogs	Recipient supposedly visited restricted website and should check which website has triggered this notification. It could be a mistake, otherwise disciplinary action may be taken.	difficult
HR System	Company converts to a new HR System with many benefits. Account needs to be activated, otherwise important information might be missing.	very difficult (without phish scale)
Microsoft	Office license expires and needs to be renewed.	difficult (without phish scale)
Virus	Device is infected with malware and will be destroyed in approx. 15 minutes. Cleaner.exe in attachment to remove malware.	medium
VR Bank	Call for an obligatory security check of account at German VR-Bank. Otherwise liable in case of misuse.	easy to medium
PayPal	Account is restricted due to unauthorised access. User needs to verify Data.	easy to medium
CEO	CEO is in call and has urgent task that needs to be done.	very difficult (without phish scale)
Password Phishing	Password is expired and needs to be changed. Otherwise account will be blocked.	medium to difficult
Scanner	Scanner sent a Word file in attachment to the recipient.	difficult
Girl	Unknown girl sends zip file with title "hot pictures"	easy
DHL delivery	Delivery is stuck at customs due to missing fees that still have to be paid.	easy

Table 2: A description of all legitimate e-mails we created for our survey.

Legitimate E-mail	Description
Doodle	Doodle survey for new time slot for weekly team meeting from supervisor.
Job Application	Job Application on a XING job offer as an IT Consultant. PDF with application in attachment.
Feedback	Google Forms survey about feedback on company event.
Resilience Training	Invitation to a training course about resilience. Registration via bit.ly link.
Report	CEO sends annual report about first quarter of the year. PDF in attachment.
Dropbox	Colleague from marketing shares photos from company event via dropbox.
Attachment	Colleague from financial department asks to check a receipt in attachment.
Passwort Legitimate	In English language. Notification that password needs to be changed in next days.
Present	Collection for birthday present for supervisor from colleague.
Office Moving	Information about new office allocation and excel sheet in attachment with exact information from facility management.
Heating	Information about short heating outages due to maintenance from facility management.
IT-Service	IT-Service informs that outlook briefly not available due to maintenance work.
Staircase	Facility management informs about closed staircase due to construction work.
Cake	Colleague has leftover cake placed in kitchen.
Callback	Applicant request an urgent recall. Number in PDF in attachment.

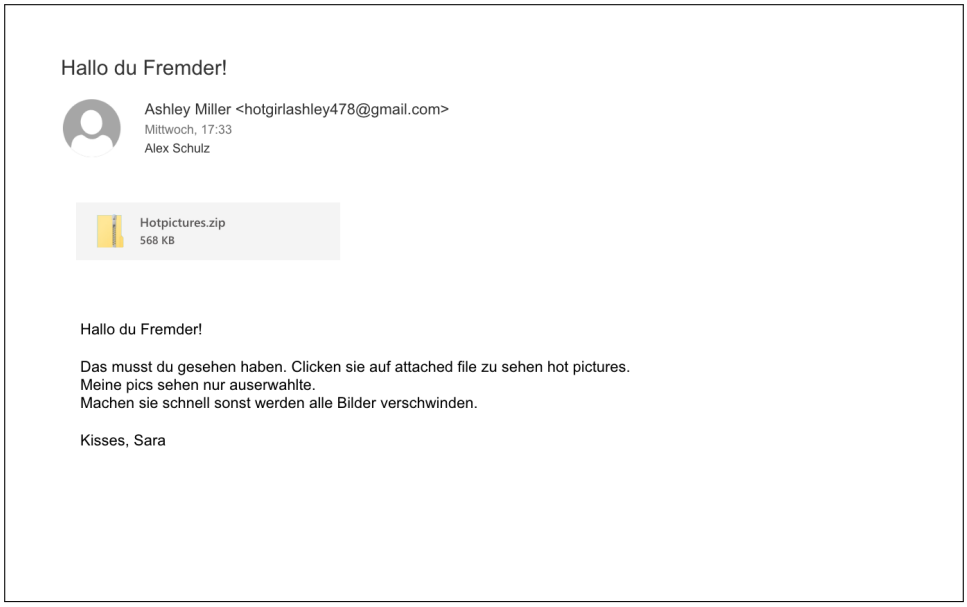


Figure 4: Phishing e-mail - Girl

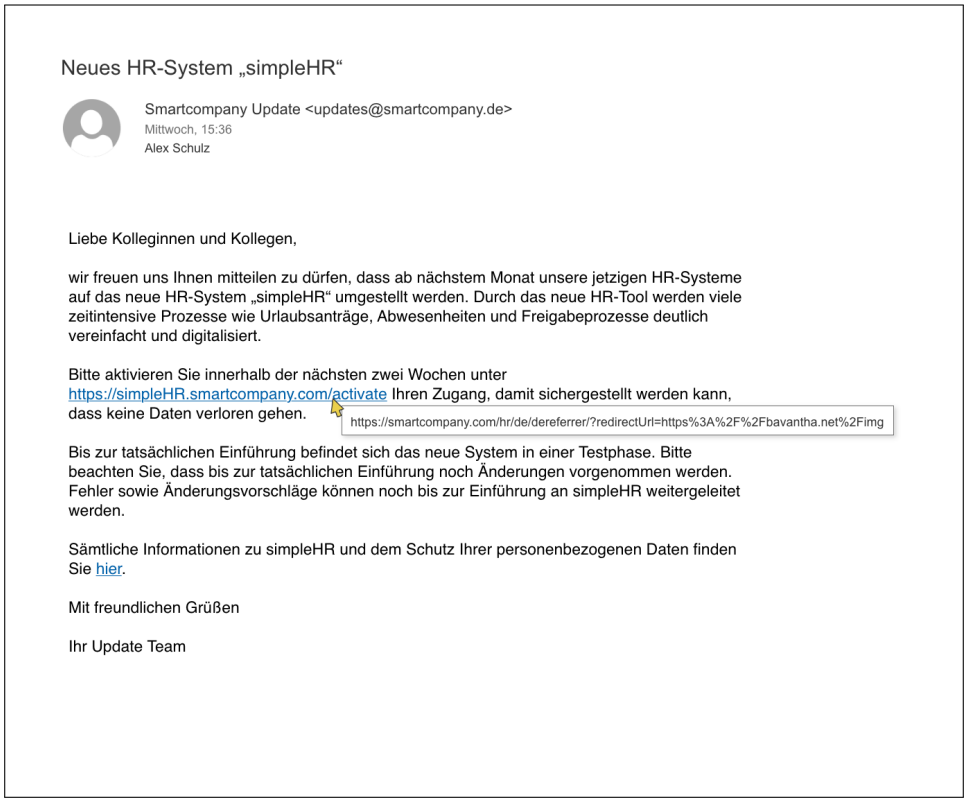


Figure 5: Phishing e-mail - HR System

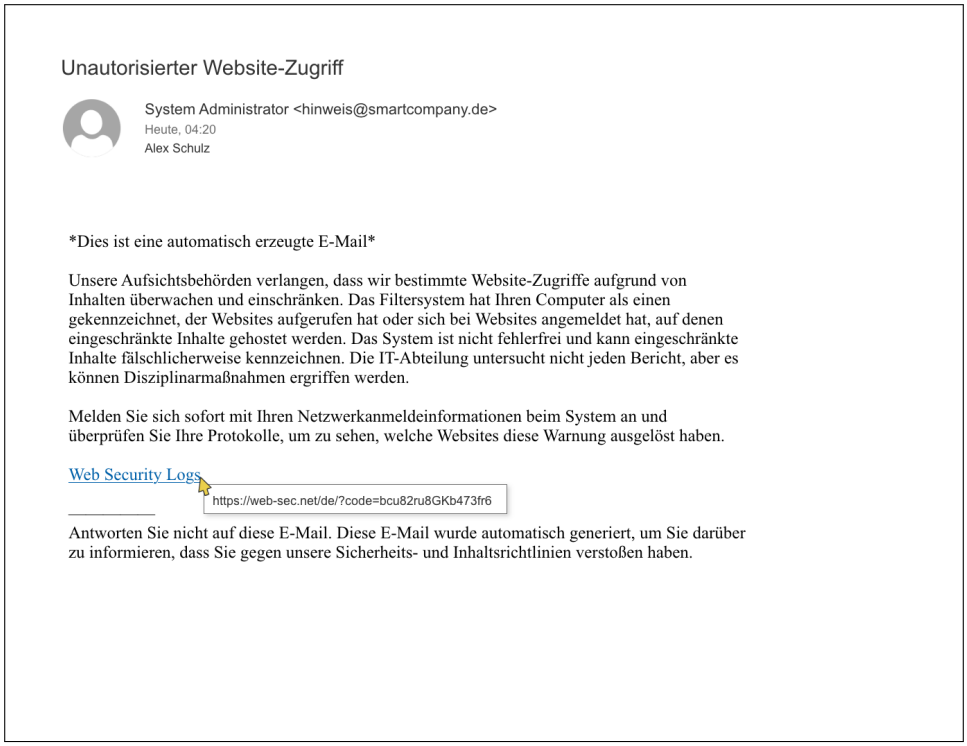


Figure 6: Phishing e-mail - Weblogs

Figure 7: Phishing e-mail - Scanner

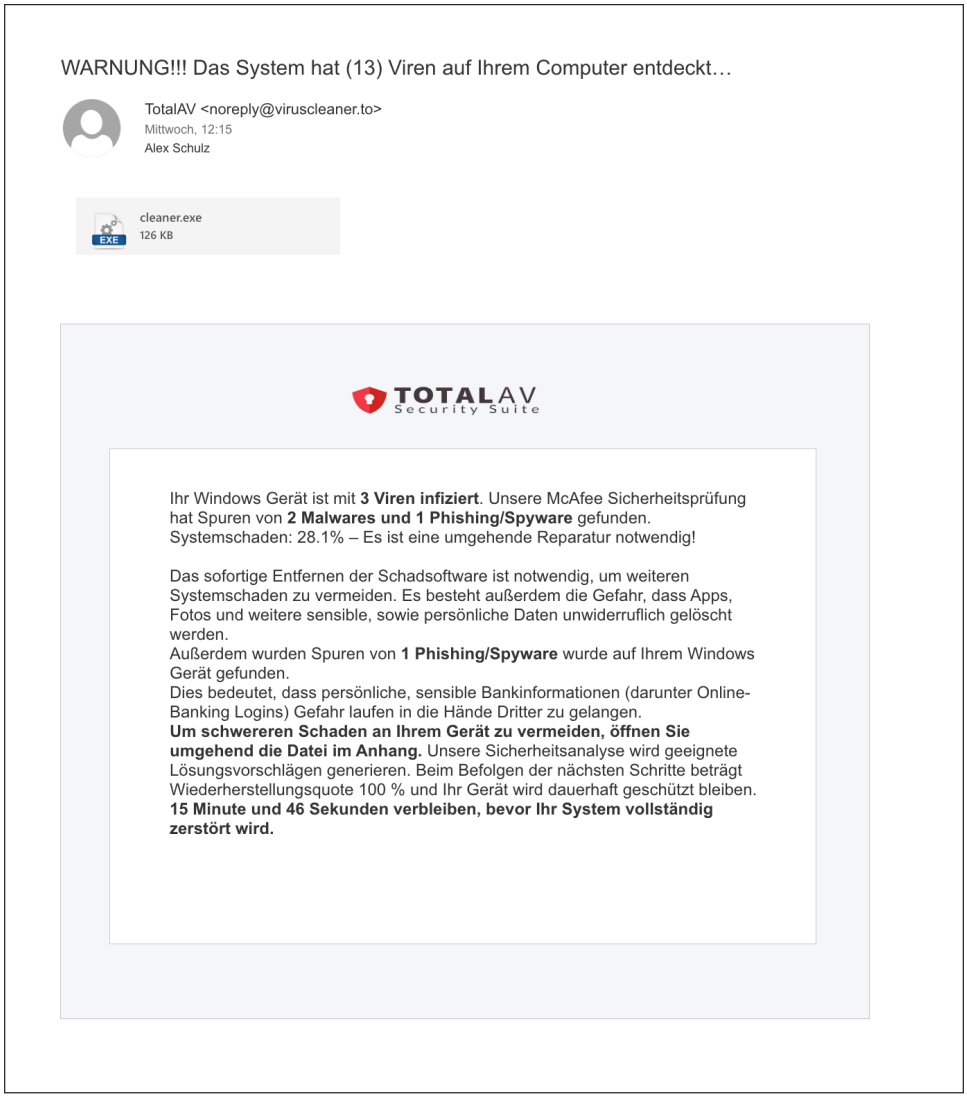


Figure 8: Phishing e-mail - Virus

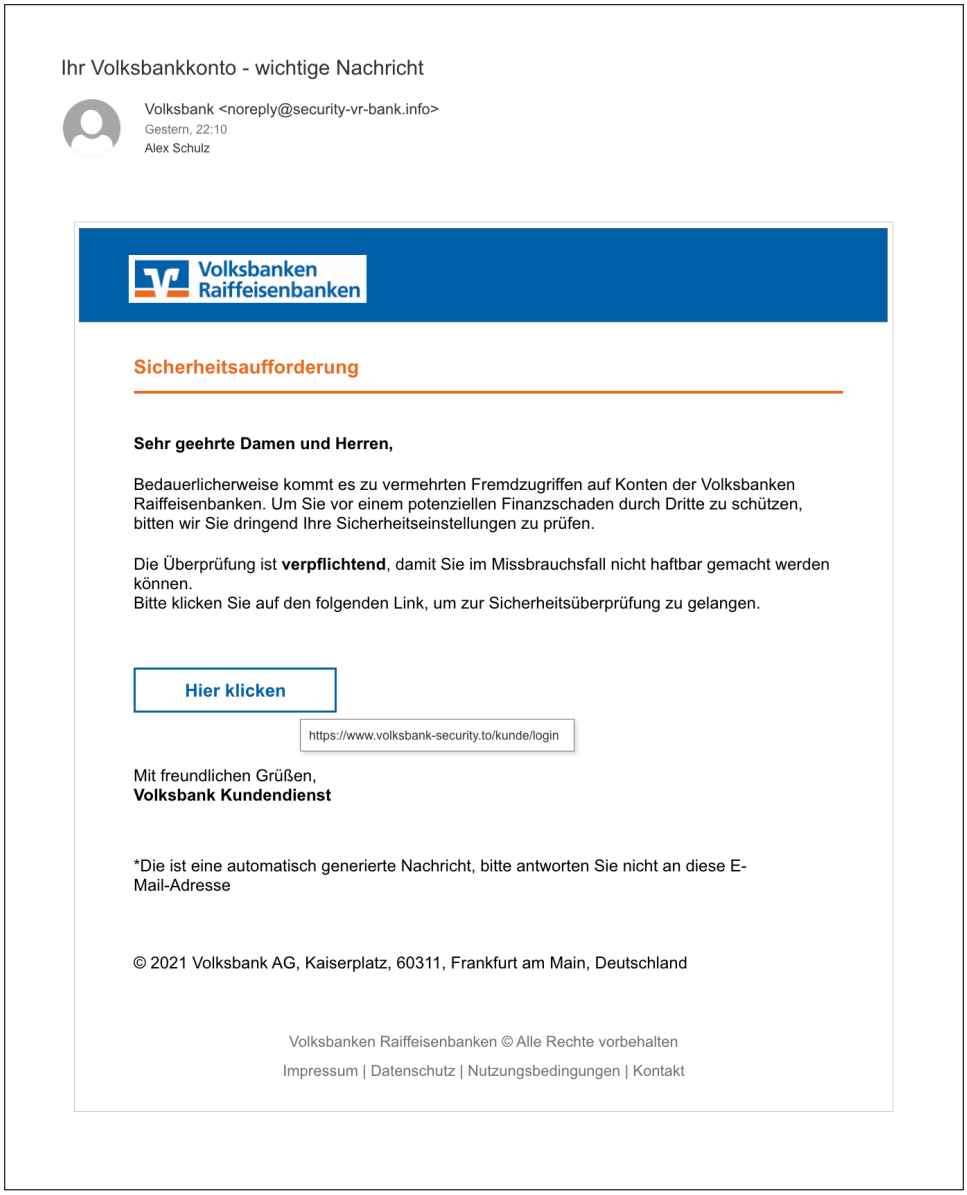


Figure 9: Phishing e-mail - VR bank

Figure 10: Phishing e-mail - CEO

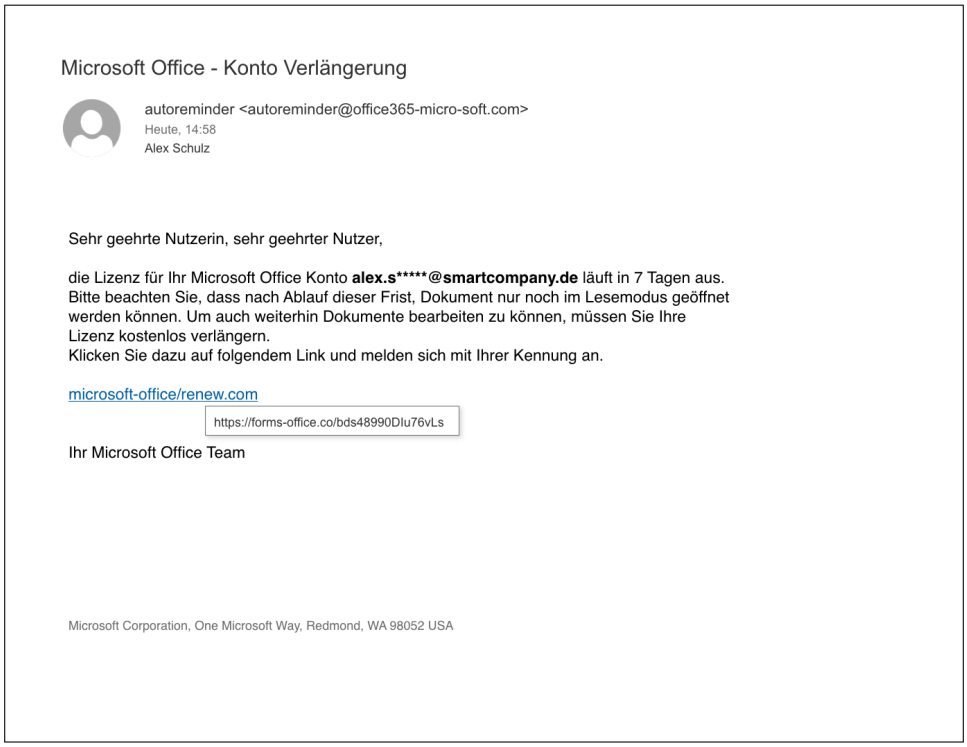


Figure 11: Phishing e-mail - Microsoft

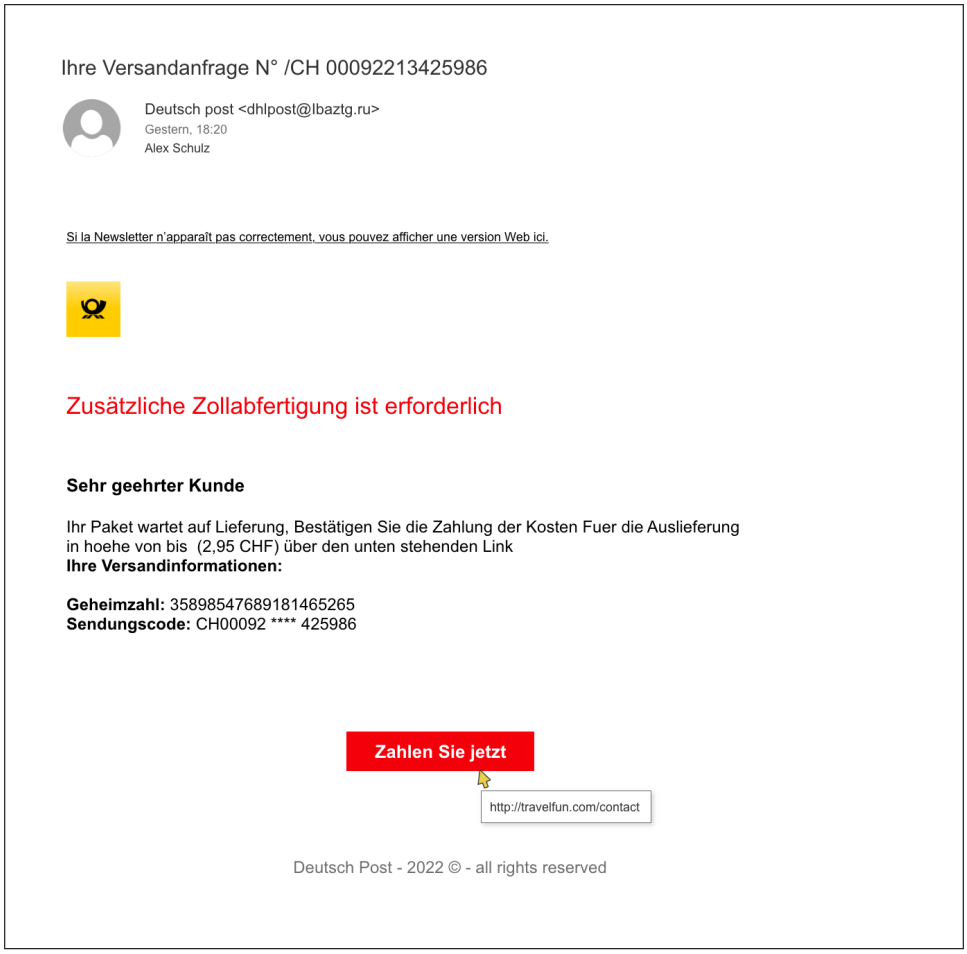


Figure 12: Phishing e-mail - DHL delivery

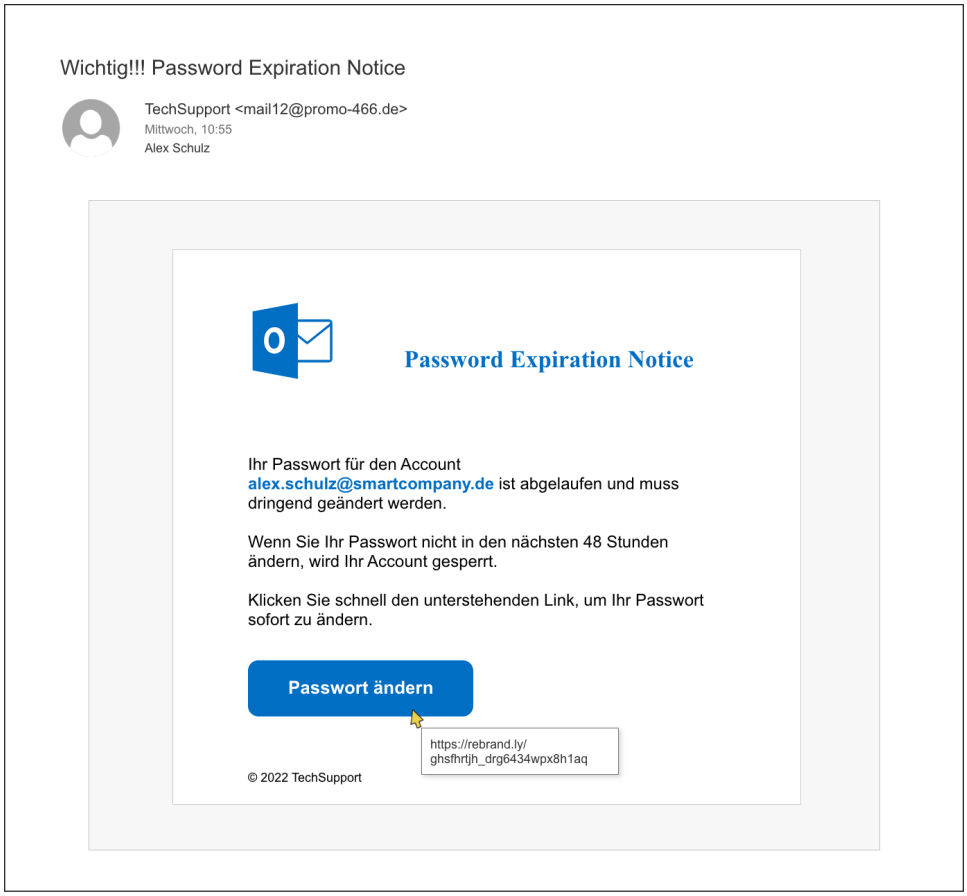


Figure 13: Phishing e-mail - Password

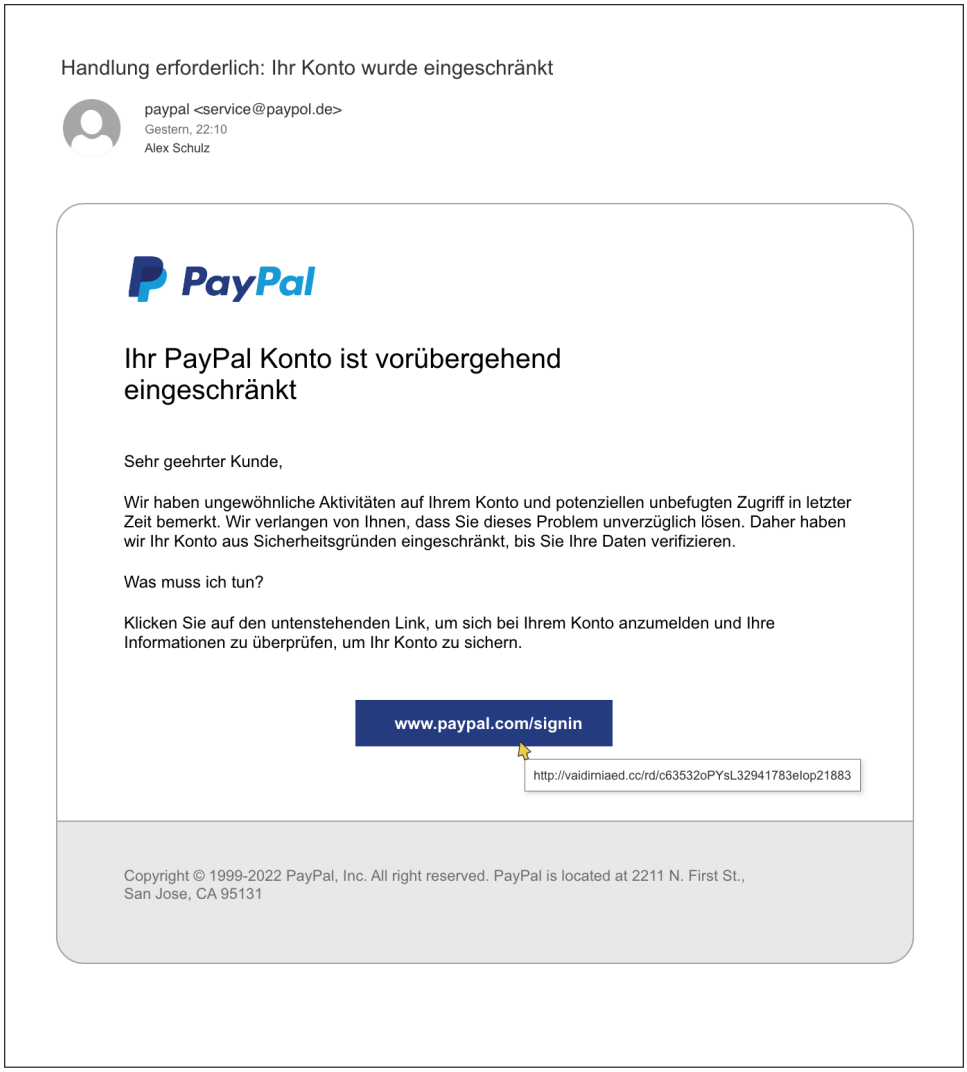


Figure 14: Phishing e-mail - PayPal