

# How Users Investigate Phishing Emails that Lack Traditional Phishing Cues

Daniel Köhler<sup>[0000–0003–3121–3888]</sup>, Wenzel Pünter<sup>[0000–0002–8218–0732]</sup>, and  
Christoph Meinel

Hasso Plattner Institute  
University of Potsdam, Germany  
{daniel.koehler,wenzel.puenter}@hpi.de

**Abstract.** Phishing is still one of the prevalent threats targeting private persons and organizations. Current teaching best practices often advocate cue-based investigation methods. Previous research primarily confronted participants with phishing emails showing such indicators to assess the success of different education measures. Our large-scale mixed-methods study challenges the behavior of 4,729 participants with four phishing emails that lack technical cues. The phishing emails concerned entirely fictitious entities and were directed at participants in their private lives, recruited from the online education platform *openHPI*. For our analysis, we apply the human-in-the-loop model for interaction with phishing content to investigate participant behavior when their learned best practices for detection fail. The primary indicator of enhanced phishing resiliency observed in our study was awareness of missing context to the supposed entity. Such context is often successfully enhanced by web searches, significantly contributing to decreased phishing susceptibility.

**Keywords:** Phishing Investigation, Cybersecurity Awareness, User Study

## 1 Introduction

Phishing, social engineering delivered via emails and other communication channels [29], has been the primary initial access vector used by cyber threat actors in 2022 [6]. In phishing campaigns, the adversary often tries to trick users into entering sensitive information on a malicious website [29] or to lure the user into performing a self-harming action [7]. This goal is often achieved by impersonating a legitimate third-party entity known to the target and counterfeiting its website and branding.

Due to the high practical relevance of this threat vector, fellow researchers have published numerous works on technical and human aspects of phishing in the past. Examples include aspects of phishing emails that drive their persuasiveness [26,21], such as logos and images [37]. Other research on phishing has investigated how socio-demographic features of targets impact their susceptibility [13,11,28], or how technical measures such as highlighting external emails enhance protection [39]. Traditional phishing education often covers technical or

psychological cues and triggers used inside phishing emails, such as typosquatting, to sensitize users for these indicators. In professional contexts, such education is often performed using embedded phishing training programs [1,25]. For laypersons outside professional contexts, central (e.g., governmental) institutions attempt to provide cybersecurity awareness programs by similarly highlighting common cues to identify phishing [12,2,3]. With a population primed to expect and suspect learned cues and technical features of phishing emails, such as manipulated email senders and links, we investigate the following research question, to the best of our knowledge never explicitly studied before:

*Research Question* How do people investigate phishing emails that lack traditional (technical) cues for phishing?

We performed a mixed-methods study combining quantitative results from a phishing study with qualitative results obtained in a post-study survey. We designed the phishing study according to *Staged Innovation Design*, allowing us to introduce new participants and thereby study an unbiased group of participants in each of our four interventions. We studied a total of 4,729 participants in overly private contexts recruited from the online education platform *openHPI*, to which we sent more than 14,000 phishing emails, all concerning entirely fictitious entities, without technical indicators for phishing, such as manipulated email headers, links using typosquatting, or impersonation of other companies. We obtained quantitative insights into the target variables of *link click* and *data submitted*. In a separate publication, we investigated the quantitative results of participant’s socio-demographic features towards the target variables, identifying that male participants, who are particularly young or old and of lower levels of education, are more susceptible to falling victim to phishing attacks [16].

To achieve additional qualitative insights into participants’ investigation processes when challenged with the emails, we collected survey answers from 950 participants. We map participants’ investigation approaches to the human-in-the-loop (HITL) model, which describes the process people follow for phishing investigation, as systematized in previous studies with smaller participant groups [33,35,20]. During our analysis in Section 6, we touch on human interaction with phishing content, from investigative approaches to time spent on web pages. We thereby foster three main contributions to the body of research:

*Contribution 1* We map survey responses from a large-scale real-world phishing study to the HITL model contributing to the systematization of human phishing investigation behavior. We present three resulting taxonomies in Section 6.1.

*Contribution 2* We observe that identification of (missing) context during the phases of *Expect* and *Suspect* in the HITL model significantly decreases *link click* and *data submission* rates of participants (cf. Section 7.1).

*Contribution 3* In Section 7.2, we identify and discuss that web searches used to generate more context on the entity or topic posed in the email significantly helped participants identify our emails as phishing.

## 2 Background

*Phishing* is a form of social engineering that can be modeled as a cycle of interactions between attacker and victim, influencing a victim’s trust and subsequent actions [18]. Victims receive and assess phishing emails. Upon following phishing links and visiting web pages, they face new, convincing information from attackers, which they must contextualize to decide how to act. These interactions can be described using a human-in-the-loop model [5]. *Wash* and *Nthala* identified a process that both experts [33] and non-experts [35,20] follow when investigating a piece of (phishing) content, deriving the HITL model for phishing email investigation. It consists of the following steps:

1. **Noticing** While viewing a piece of content (e.g., mail, website), humans extract features like the type of email, context, sender, layout, or URL format.
2. **Expecting** People subconsciously compare the noticed content features to their expectations.
3. **Suspecting** When features deviate from the expectation or trigger a learned cue, suspiciousness emerges.
4. **Investigating** When suspiciousness is raised, people begin with investigative behavior like hovering over or following links, reading the imprint, or contacting the sender.
5. **Deciding** Based on the investigation, humans decide how to interact with the content or collect more info.
6. **Acting** Depending on the result, people might respond differently to a message (e.g., continue, delete, or ignore).

## 3 Related Work

Fellow researchers have already studied various parts of the phishing landscape. Previous studies on the effectiveness of cybersecurity awareness education, particularly phishing, usually focused on highlighting technical cues to identify maliciousness [26,4]. As such, fellow researchers have explored URLs [36,8,27], spelling and grammar [10,22], visual cues such as images and logos [22,10,21] as only few of the core criteria used to assess phishing emails.

Other researchers have evaluated approaches such as story-based education [34], which still used technical best practices such as “*Hover over a link to see where it really goes to*” as taken from *Wash and Cooper’s* 2018 study [34]. Contrasting, *Jensen et al.* have started to evaluate approaches to enhance mindfulness during email analysis [14]. Mindfulness in the respective study was triggered through considerations of the context of the email, such as “*Why would the sender need me to do this?*”. The authors identify that while mindfulness approaches help participants with high email skills, they cannot replace cue-based approaches like those highlighted previously.

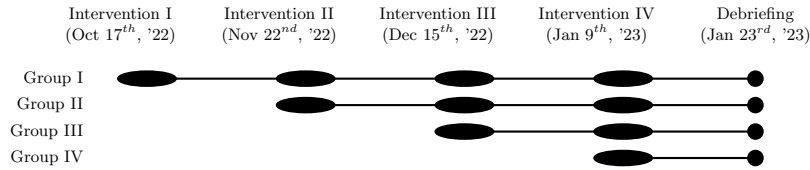
Mindfulness training, as used by *Nguyen et al.* [19] and *Jensen et al.* [14], supports the assumption that more strongly assessing email context should be

the key to achieving more resilience against phishing attacks. Still, many educational programs we observed in the wild focus on rule-based phishing training. Such educational programs attempt to provide users with rules (e.g., *Check for the spelling of the domain*) to educate users on what they should be looking out for. However, to the best of our knowledge, analyzing emails without cues for phishing is missing from previous research.

*Wash* and *Nthala* have previously worked on the human-in-the-loop model, which we use as a foundation for our study. *Wash* initially interviewed 21 experts on instances where they successfully identified phishing attacks in an exploratory study [33]. Based on the HITL model, he derived a process that expert users follow to identify and assess phishing emails. Building on this process, *Wash*, *Nthala*, and *Rader* surveyed 297 non-experts from the US on their experiences with phishing emails [35], identifying that the investigation process of non-expert users is similar to the process of experts. *Nthala* and *Wash* verified the previous findings in a study with 31 non-expert users, sending them a phishing message and interviewing them on their experience with the email [20]. They highlight that non-expert users often depend on their social connections, unlike expert users who primarily rely on technical investigations.

The previous studies on HITL models relied on interviews. Hence, the number of participants was limited. In our study, we expand the number of studied users, further challenging them with a larger variety of emails, thereby contributing to contextualizing real-world phishing investigation processes. As we track participants’ actual behavior, our study allows quantifying the success of individual investigation measures mentioned by the participants.

## 4 Method and Study Design



**Fig. 1.** Overview of our *Staged Innovation Design* study and timeline in which we ran it. *Debriefing* is explained in Section 4.2.

Our study’s quantitative-qualitative mixed-methods approach combines a large-scale field study with a survey for qualitative reflection of participants’ behavior during the study. The study has been designed according to *Staged Innovation Design* [32], whereby we introduce participant groups into the study across four different phishing interventions. During each intervention, participants receive a phishing email containing a link to a phishing website asking

for personal data and credentials. For the analysis, we aggregated participant behavior across the emails. We use the terminology of *intervention* taken from educational research referring to the times in which researchers interact with study participants, such as sending a phishing email in our case. The interventions were distributed in approximately four-week intervals (cf. Figure 1) to exceed knowledge retention periods reported in previous studies [15]. One week after the last intervention, all participants received a debriefing email and were asked to participate in the optional post-study survey. The study targeted German-speaking participants.

#### 4.1 Participant Recruitment

Participants were recruited from *openHPI*, an online education platform with over 100.000 registered users [17], where university lecturers provide free online courses on IT-related topics to the general public. Recruitment happened in the form of an additional consent available on the platform covering “Research at HPI”, which learners of the platform could provide in their profile. The consent covered data processing and analysis outside the education platform and email communication in the context of research studies. Upon logging in to the platform during the study duration, learners received a one-time notification that the new consent was available. The default value for the consent was *off*, requiring users to actively opt-in to the study. Then, they could opt-in and -out of the study at any time during the study period. Studying real-world interactions with phishing content poses the challenge that many participants will never open phishing emails. Therefore, we did not set an upper limit on the number of participants for our study but included everyone providing consent during the study period.

#### 4.2 Ethical Study Design

Studying human subjects requires researchers to closely consider ethical questions, such as the mental load on participants. Therefore, human subject studies should generally be assessed by an Institutional Review Board (IRB) and require consent from the subjects to be studied. Consenting to a study generally means that individuals who are fully informed about it actively agree to participate [24]. In cybersecurity, particularly phishing research, participants could provide informed consent, e.g., at the beginning of a lab study. However, providing all information on a study can lead to biases of the subject, altering the study results [9]. Instead, researchers can use deception by withholding essential information on the study design from participants [9]. Deception is only deemed an option if the study is of minimal risk to the participants and requires the researcher to *debrief* the participants upon completion, i.e., provide all previously withheld information.

Our study design uses deception upon participant recruitment. We did not inform consenting participants that we would perform a phishing study. The IRB of the University of Potsdam and the data protection officers of the conducting

institute approved this study design. When visiting the online education platform *openHPI* starting in September 2022, learners could provide and revoke consent to receive email communication for research projects. Once they provided their consent, they would be included in every following iteration of our study. E.g., we included a learner who provided the consent on Dec 12<sup>th</sup> for the third and fourth iteration (cf. timeline in Figure 1).

Phishing attacks are no unusual threat for any user of the Web and Internet. Hence, during the assessment of our study, the IRB agreed that the planned research, including deception to retrieve consent, poses minimal risks to participants. For our web pages, we ensured that no personal data, such as usernames, passwords, or address information, entered by the participants would be transmitted to our servers. All other data, e.g., reaction to an email and behavior on the web page, was collected in pseudonymized form, i.e., it was only labeled with a pseudonymized identifier, not a user’s email address or username. All users were *debriefed* with the final email in January 2023, a week after intervention four of the study had been sent to all participants<sup>1</sup>. To ensure the debriefing email reached all participants, we sent it using the official email servers of *openHPI*. This provided a trust anchor for the users and ensured that the potential lack of reputation of our phishing domains did not limit email delivery. That debriefing email contained all information on the study, the researchers involved, consent and legal information, a link to the survey, and a link to revoke the provided consent. The user data was removed before further analysis if the consent was revoked. Twenty-one users revoked their consent throughout the study.

To keep the mental load on participants as low as possible, we included *debriefing information* in all our resources to be found whenever an in-depth investigation would be performed. Such were, e.g., hidden as white text inside the email, the web page’s source code, and the web page imprint. Further, once participants, e.g., replied to the emails or contacted the supposed support email addresses, we also debriefed them. The debriefing contained the scientific and legal background of the study and information on how to resign from the study by withdrawing consent. The multi-staged nature of our study design poses the challenge to monitor when a participant has been debriefed and should be excluded from further analysis. We discuss the challenges arising from the debriefing of users in Section 7.3.




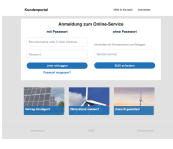


### 4.3 Email and Webpage Content Design

Figure 2 presents screenshots from all four emails and web pages sent to participants. Content and entities used across all emails were fictitious and created solely for this study. We prepared an email and webpage for each study intervention, which was designed based on real-world designs of similar companies. While similar in design to known companies, none of our emails featured traditionally taught technical cues of phishing emails, such as typosquatting in

<sup>1</sup> Emails were sent and delivered to all participants throughout approximately one week for each iteration. This measure ensured that no sudden traffic spike from formerly little-known domains would put the respective domains on a spam list.

links. To maximize data on email and webpage interaction by users collected in this study, we designed persuasive emails, basing various design decisions on the more convincing vectors identified in previous research. The emails relied on the more significant psychological vectors such as time pressure, trust, or financial loss [23,37]. To incorporate these vectors, we chose fitting topics for the emails: supposed package deliveries (emails I and IV), a mail from an energy company concerning rising energy prices (emails II), and a supposed payment confirmation (email III). Emails III and IV used a personal salutation to enhance the email delivery rates, while emails I and II relied on a generic salutation. This change was introduced to counteract emails being blocked as spam and is visible in a change of delivery rates as depicted in Table 1. Throughout all emails, we omitted any traditional cues of phishing emails, such as spelling mistakes. Further, we tried to include images or logos in each email to enhance persuasiveness [37].

Upon following a link from our phishing emails, users were presented with company web pages. These continued the email’s theme and topic, persuading the users to enter personal information such as an address, username, or password. Besides, for example, a package tracking website as a landing page for emails I and IV, each domain hosted further web pages, such as the home page and an imprint of the fictional company.

Entity	Email I	Email II	Email III	Email IV
	<b>paket-info.org</b>	<b>verbraucherschutz-strom.de</b>	<b>pay-online.at</b>	<b>easy-paket.eu</b>
Email				
				

**Fig. 2.** Phishing content sent to the participants in the four iterations, ordered left to right. Large-scale images are available in the Appendix, Figure 8, Section B.

The debriefing email informed participants of the nature of the study that had been conducted (cf. Sec. 4.2) and invited them to participate in our survey. To not impact the participants’ alertness and, thereby, future reactions to our emails, the debriefing and survey could only be performed after the completion of the entire study. Due to the post-hoc nature of the survey, sent 3.5 months after the initial phishing emails, we expect some inaccuracies in participants’ memories to occur. Such could be that participants only remember particularly important or surprising aspects of the emails [21]. We discuss this limitation in Section

8. In the survey (cf. Appendix, Section A), we retrieved (socio-) demographic information on the participants and their perception of the emails and web pages. Throughout the survey questions, participants could select pre-provided answers from multiple-choice lists and additionally provide free-text answers. For evaluation, *Wenzel Pünter*, second author of this manuscript, manually coded the free-text responses and mapped all responses to the HITL model.

To ensure an email contained no technical cues for phishing, we purchased all four domains, registered them with a new public IP address, and sent emails directly from these domains. As no impersonation was performed, links in the emails did not rely on special characters or typosquatting to counterfeit a third-party entity. In order to enhance our delivery rates, we sent a few hundred emails from the new domains to our inboxes at various webmail providers before sending the actual study emails. This preparation reduced the number of our emails being rejected by webmail providers. Across the interventions, we were able to improve our delivery rate to 99.70%<sup>2</sup> in the final iteration (cf. Table 1).

#### 4.4 Data Collection & Cleaning

We obtained two datasets: the survey responses and the tracking data from the phishing campaign with four iterations. The tracking data covered all four stages of the funnel of phishing email interaction as described by the following actions:

1. **Delivery** Emails have been sent using a commercial gateway. The time of email delivery has been recorded for each email.
2. **Open** Dynamic elements tracked when the email was opened.
3. **Clicking** on a phishing link from the emails opened the website using HTTPS.
4. **Submit** The phishing website requested to enter personal data (username and password, or address), revealing a *debriefing* page upon submission.

The server recorded the timestamp, requested page, IP address, and user agent for the click and submit stages. The websites contained JavaScript-based tracking elements that allowed recording the load and unload timestamps, screen and window dimensions, scroll, mouse, and touch behavior, window visibility changes, input blurring, and focus events. The data recorded by the server, in particular, served as ground truth for the following analysis of user behavior on the web pages. To prepare the analysis, we performed data cleaning of the tracking data from our study. We reduced the web page requests to our servers which we included for further analysis from 6,881 to 5,275:

1. To analyze actual user behavior, automated mail sandbox and security system traffic were cleaned using ASN- and user-agent-based filtering. We excluded traffic from networks of carriers, hosting- and security service providers, as well as requests issued by non-browsers, such as bots, previews (iMessage or Discord), or headless browsers to remove automated reactions to our emails, ensuring our data contains only, e.g., link clicks by human users.

<sup>2</sup> This delivery rate is based on email server acceptance. Email classification, e.g., into the *junk/spam* folder as done by secondary filters, can not be tracked in our setup.



2. Automated traffic from commercial IP spaces was identified using a commercial dataset on IP addresses, allowing the filtering for traffic from data centers, VPNs, and anonymization services. Similar to (1.), we excluded highly-similar requests from commercial data centers assuming automated behavior.
3. *Debriefed* participants were accounted for by excluding any activity after a participant was exposed to a disclaimer on our web pages.

## 5 Overview of Study Data and Participant Population

The study included 4,729 participants, to which we sent 14,123 phishing emails. Of these emails, 6,027 (42.68%) have been opened, whereby 1,549 users (32.76%) clicked on the contained phishing links and 446 of these (28.79%) submitted personal data. 950 participants (20.09%) answered the post-study survey, whereby the number of responses varies across the questions. 26 participants (0.55%) replied to at least one of the phishing emails during the study, assuming the content was legitimate. In other email responses, participants, for example, highlighted that they liked the additional learning experience and that they were adequately prepared by any of the courses they previously took.

**Table 1.** Overview of the participation rates across the intervention funnel stages.

Iteration	I		II		III		IV	
	N	Share	N	Share	N	Share	N	Share
<b>Sent</b>	1,955		3,483		4,260		4,729	
<b>Delivered</b>	1,871	95.70%	3,360	96.47%	4,177	98.05%	4,715	99.70%
<b>Opened</b>	851	43.53%	1,218	34.97%	1,844	43.29%	2,114	44.70%
<b>Clicked</b>	311	15.91%	222	6.37%	359	8.43%	657	13.89%
<b>Submitted</b>	92	4.71%	35	1.00%	51	1.20%	268	5.67%
<i>Replied</i>	0	0.00%	0	0.00%	17	0.40%	9	0.19%

Table 2 provides an overview of participant demographic information as provided during our survey. A Kolmogorov-Smirnov test of the sample data shows a significant skew compared to the distribution of sexes and age groups ( $K = 1.0000$ ,  $p = 0.0286$ ) in Germany. The most deviation in age is explained by people aged 60+, who experience a lower Internet penetration rate.

Another demographic factor considered in the study is the level of education reached by participants. The responses have been categorized according to the UNESCO ISCED-2011 taxonomy [30]. 934 participants (98.32%) have reported their highest level of education, whereby 3 (0.32%) have reached Primary education, 36 (3.85%) Lower secondary education, 163 (17.45%) Upper secondary education, and 732 (78.37%) a Bachelor’s degree or any equivalent higher level of education. Depending on their work situation, participants might have different exposure to phishing content and training. Therefore, working participants

**Table 2.** Overview of participant socio-demographic information as provided in the post-study survey. In total, 950 participants (20.09%) replied to the survey.

Feature	# Responses*	Statistics							
Gender	925 (97.37%)	Gender # Responses Share	Male	Female	Other				
			722 76.00%	195 20.53%	8 0.84%				
Age	934 (98.32%)	Age Group # Responses Share	< 20	20 – 29	30 – 39	40 – 49	50 – 59	60 – 69	> 70
			13 1.37%	59 6.21%	103 10.84%	164 17.26%	253 26.63%	209 22.00%	133 14.00%
Level of Education	934 (98.32%)	Degree of Education # Responses Share	Primary	Lower Secondary	Upper Secondary	Bachelor's	Master's	Doctoral	
			3 0.32%	36 3.85%	163 17.45%	157 16.81%	483 51.71%	92 9.85%	
IT Usage	Work: 773 (81.37%) Home: 869 (91.47%)	Usage	Always	Daily	Regularly	Rarely	Sporadically	Never	
		# Responses	425	262	38	11	7	30	
		Share	54.98%	33.89%	4.92%	1.42%	0.91%	3.88%	
		# Responses	283	490	81	4	9	2	
Work Industry †	811 (85.37%)	Industry Code # Responses Share	H	J	K	O	P	Q	other
			39 4.81%	478 58.94%	45 5.55%	124 15.29%	193 23.80%	53 6.54%	94 11.59%

† Industry Codes according to UN ISIC Rev. 4 [31], e.g. **H**: Transportation and Storage, **J**: Information and Communication, **K**: Financial, **O**: Public Administration, **P**: Education, **Q**: Health and Social Work

were asked to state the industry they are currently working in. The responses were classified according to the UN ISIC Rev. 4 primary industry groups [31]. 811 (85.37%) working participants stated their industry in the survey, whereby 654 (68.84%) participants associated themselves with only one industry and 157 (16.53%) mentioned multiple. 58.94% of our participants associated themselves with the *Information and Communication* industry, followed by 23.80% in *Education* and 15.29% in *Public Administration and Defense*.

The earlier outlined distribution of participants shows a bias of the study sample compared to the general population of Germany concerning sex, age, level of education, and work industry. The studied population is overly male, has not reached the age group 60+, has an above-average education level, and primarily works in information technology, education, and the public sector. We surveyed that most participants use IT at least daily in both work (88.87%) and private (88.96%) contexts, thereby judging that we observe a group with a high affinity towards IT systems. We discuss the following two derived biases in our population in Section 7.3.

**Bias 1** We observe the foremost discriminator from the average population and, thereby, potential bias to our study to be the overly technical population.

**Bias 2** All participants were recruited from the online education platform *openHPI*, which offers particularly IT education. Therefore, our participants will likely be more interested in IT methods, tools, and technology.

## 6 Study Results

During our evaluation, we mapped participant responses from the survey to the different phases of the human-in-the-loop model introduced earlier. To ensure

unbiased participant responses, we formulated our survey questions as broadly as possible (cf. Appendix, Section A). We applied the classification by manually labeling participants’ (free-text) responses and mapping them to the different stages of the HITL model based on the reported actions. The following sections present the results of our classification.

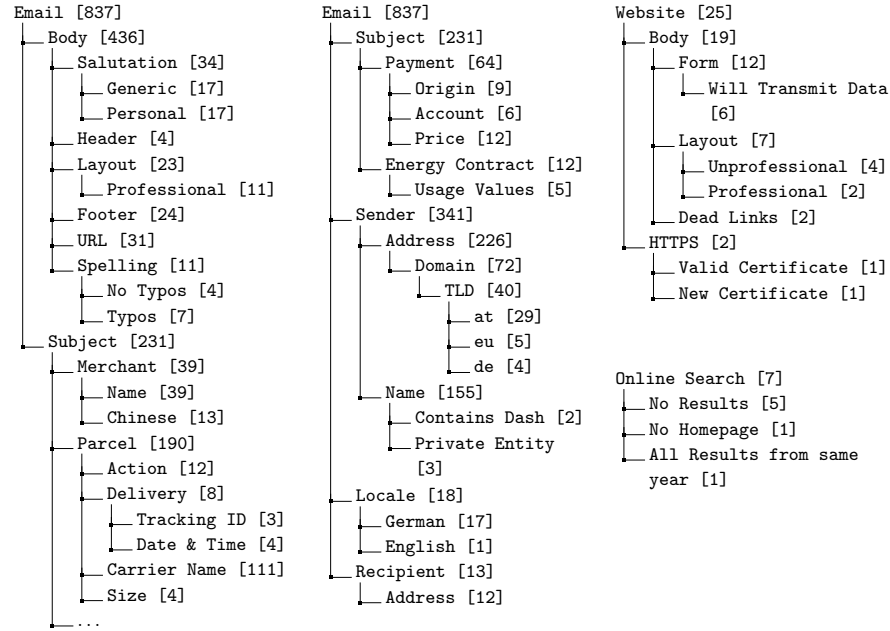
### 6.1 Mapping of Responses to the Human-In-The-Loop-Model

*Contribution 1* We contribute to the systematization of phishing investigation behavior by coding and mapping responses from participants of a large-scale study in overly private contexts to the HITL model.

As part of the survey, participants were asked to explain their actions for each phishing email received across the four iterations. Properties mentioned in the responses were classified according to the human-in-the-loop model introduced in Section 2 and manually clustered hierarchically. This section provides an overview of the explored answer space.

This section is structured alongside the human-in-the-loop model presented by Wash *et al.* [35], with the phases of *Notice*, *Expect*, *Suspect*, *Investigate*, *Decide* and *Act*. However, our study methodology partly limits the exact assignment of an answer to a precise stage in the model. For example, we asked participants which aspects of the mail caused their suspicion (*Suspect*). These differ from participants’ expectations (*Expect*). Due to our study setup (*post-study questionnaire*), we could not interview participants on their *actual* expectations for, e.g., package delivery emails before sending our study emails and survey. Still, some answers did provide information on the participant’s expectations, such as P101: “*Layout of the mail did not correspond to, e.g., UPS, DPD, etc.*”, which provides us with the information that the participant would *Expect* a package delivery announcement via email to look like the ones they are used to. The answer, however, was provided because the layout of our email triggered the participant’s suspicion. Therefore, we evaluate the two stages *Expect* and *Suspect* alongside each other.

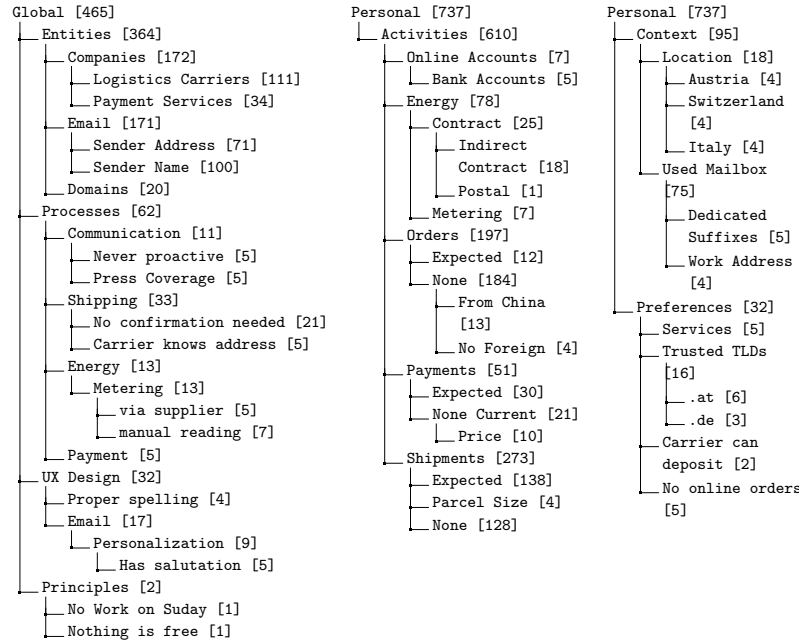
**HITL-Model: Notice** Figure 3 summarizes the features that were *noticed* by participants. We manually mapped all free text answers to the respective questions (cf. Appendix, Section A.2: Q10, Q12) and structured them hierarchically. The observed features center around metadata, content, and context of the received emails, seen phishing websites, and conducted online searches to investigate the legitimacy of content. For example, 17 participants noticed that the *Email Salutation* (*Email* ► *Body* ► *Salutation*) was very *Generic*. In contrast, 17 [others] noted the *Personal* salutation we employed in emails III and IV. Similarly, a total of 40 participants noticed the *Email Sender’s TLD* (*Email* ► *Sender* ► *Address* ► *Domain* ► *TLD*), for which some took particular notice of, e.g., *.at*, or *.eu* ( $N_{.at} = 29$ ,  $N_{.eu} = 5$ ). However, few observations do not match reality, only participants’ perceptions, biased by imperfect memory. E.g., we sent all content in German but never in English, how some participants reported to have observed it (cf. Appendix, Figure 9, *Email* ► *Locale*).



**Fig. 3.** Highlights of hierarchically structured *Noticed* properties named in survey responses. Participants *noticed* aspects within the **Email**, **Website** and during their **Online Searches**. [Numbers in brackets] refer to the count of mentions. Figure 9 in the Appendix, Section C shows the entire figure.

**HITL-Model: Expect & Suspect** Properties that were *suspected* and thereby differ from what was *expected* by participants split into a context that is unique to the person itself (**Personal**) and expectations that emerge from a person’s assumptions about the world and its relationships (**Global**). An overview of suspected and expected aspects mentioned in the survey is provided in Figure 4. For example, 111 participants mentioned global expectations towards emails by *Logistics Carriers* (*Global* ► *Entities* ► *Companies*). Regarding *Personal* expectations, 12 participants reported to have had expected orders (*Personal* ► *Activities* ► *Orders*). In comparison, 184 participants claimed they were not expecting orders or, e.g., never ordered from foreign countries ( $N = 4$ ).

Based on their observations and expectations, participants attempted to identify the context of the email. Often, they expressed either an event that required legitimate communication or different kinds of fraud, sometimes resulting from a data breach. Those participants who suspected fraud assumed their identity data was leaked from a service provider ( $N=20$ ), phishing ( $N=2$ ), spoofing ( $N=1$ ), or domain-specific types of fraud associated with the message content like a fraudulent order ( $N=2$ ), the abuse of a credit card number ( $N=6$ ), or a compromised PayPal account ( $N=1$ ). Typically, those who assumed legitimate communication

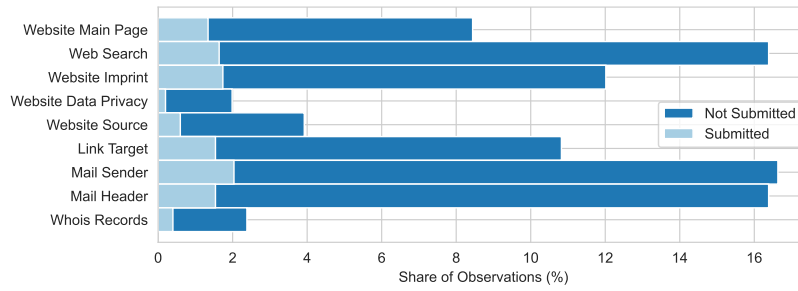


**Fig. 4.** Highlights of hierarchically structured *Expected* and *Suspected* properties named in survey responses. We differentiate between **Global** expectations that could be identical across participants and **Personal** expectations, e.g., of a concrete shipment. [Numbers in brackets] refer to the count of mentions. Figure 10 in the Appendix, Section C shows the entire figure.

suspected recent personal activities as the origin of the unwanted communication:

1. **Energy Price (Email II)** While the sent email was themed along governmental support programs in consequence of the Russian invasion of Ukraine in 2022, participants suspected legitimacy not because of this context but personal circumstances like a newly established supply contract (N=4) or the delegation of duties from the supplier (N=2) to an unknown third party.
2. **Payment (Email III)** Most of the participants who suspected the legitimacy of the payment email assumed that they missed the payment for an online order (N=13). Others believed the payment was misrouted (N=1) or the transaction was a pending refund (N=1).
3. **Logistics Services (Emails I and IV)** As with the payment message, most legitimacy assumptions centered around pending orders that participants were no longer aware of (N=22). Related events like a recent birthday (N=2), a Christmas parcel (N=1), or a current address change (N=1) can also explain the unexpected delivery message. Several participants assumed a parcel to or from another person, like their spouse (N=1), a relative (N=1), or another third party (N=1).

**HITL-Model: Investigate** Several participants mentioned how they investigated the legitimacy of the phishing content in each iteration. Investigation techniques performed by participants are - besides their representation in the HITL model - of significant interest to this study. The traditional analysis of, e.g., link targets or email headers does not provide insights in our study, as all phishing emails lacked technical cues for maliciousness. In the post-study survey, 884 participants (93.05%) reported their investigation methods. Figure 5 presents an overview of the distribution of participants’ answers on their investigation techniques, contrasting with whether they fell for any of the received phishing emails.

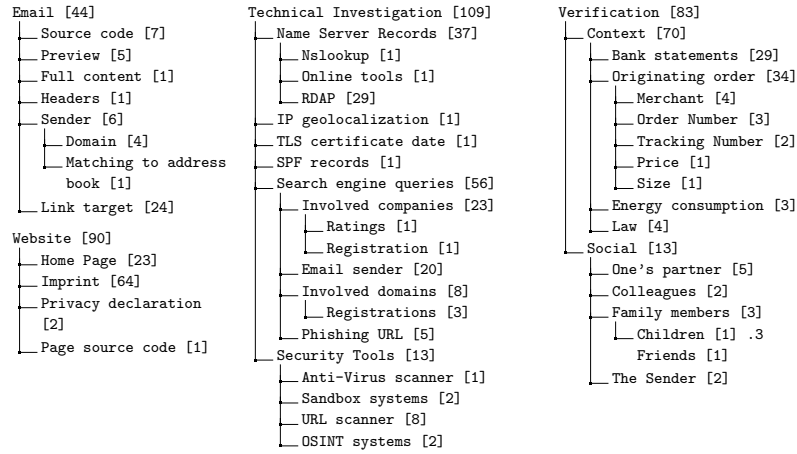


**Fig. 5.** Distribution of participants’ qualitative survey answers on investigation techniques enriched with whether they had submitted any data during the study ( $N = 844$ ).

In the free-text responses, participants mentioned that they aimed to fulfill two goals with their investigation: (a) collecting additional information and (b) verifying observations and assumptions with external information. One participant also mentioned an experiment-based approach, entering fake data and modifying URL parameters to test the web service.

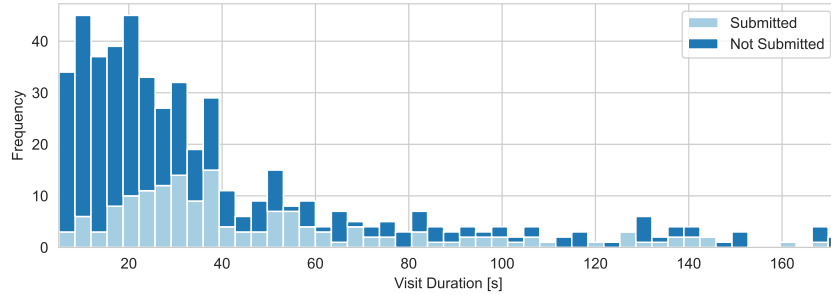
Figure 6 presents the taxonomy of participant replies. Participant investigation techniques could be grouped around the email itself, the webpage, technical investigation procedures such as investigating name server records, or verification of the supposed content of the email through either context or social contacts. Sixty-four participants mentioned having investigated the imprint of our webpage. 56 participants mentioned using search engines to investigate the supposed companies, email sender, involved domains, or phishing URLs. Various other responses covered verification of the email content; examples include verifying the context by checking bank statements or attempting to match the phishing emails to originating orders. Other participants highlight getting help from social connections such as family members or colleagues to verify the email content, which confirms the earlier introduced findings by *Nthala and Wash* [20].

Those aspects of the participants’ investigation, which targeted our phishing web pages, were measurable. As participants reported investigating the content,



**Fig. 6.** Hierarchical overview of *investigative* measures named in free text survey responses. Numbers in [brackets] refer to the count of mentions.

one could hypothesize that an increased amount of observed content reduces the chance of clicking on the phishing link or submitting data on subsequent web pages. However, content visibility and the device type (e.g., mobile or computer) did not show to be significantly correlated with data submission on the web pages. We further assessed user scroll distance and time spent with the web page as proxy factors for the investigation process.



**Fig. 7.** Histogram of time spent on the phishing page before submitting data or leaving the page. The figure shows the 90% quantile of the long-tail distribution and how many users showed the respective behavior (*Frequency*)

When users visited our web pages, we could calculate the visit duration by investigating the time between JavaScript (JS) load and unload events on the webpage. We excluded requests without those events, as the disabled JS severely limits our tracking capabilities. On the first visit, the median time spent before

leaving the webpage was 31,05 seconds. Figure 7 shows the histogram of visit durations. Tracking webpage blur events<sup>3</sup> showed that 645 participants (41.64%) left at least one of the phishing pages to other tabs or windows during their first visit to our page. We interpret this observation as a proxy factor to web research (e.g., Google search) on an entity or context provided on the webpage.

**HITL-Model: Decide & Act** The *Decide* step in the cognitive human-in-the-loop model is hardly measurable in a field study. Based on the observed, expected, and suspected properties, participants *decided* that the seen piece of content is legitimate or illegitimate and acted upon it.

In the survey, 310 participants expressed different *Actions* in response to the content in free text answers: 149 participants (48.06%) reported that they deleted the respective email, 53 moved it to the junk folder, 35 (11.29%) contacted the authors, 25 ignored the content, 24 reported it to another entity, 16 admitted clicking on the link, four blocked the sender, three replied to the message assuming it was legitimate, and one person waited for subsequent messages. Two participants mentioned that they monitored their bank accounts in the subsequent days for malicious transactions. Those participants who reported the content to a third party forwarded it to their organizational IT department (N=10), to the RDAP abuse contact (N=3), to their bank (N=1), or filed a report to their local police (N=1).

## 6.2 Impact of Features on Participants' Reactions

In the previous sections, we outlined which behavior, investigation, or observation has been reported by participants. Building on our mixed-methods approach, we have aggregated both quantitative and qualitative data. The aggregated quantitative dataset on participant behavior in reaction to the phishing emails, such as *opened the phishing email*, *clicked on links*, or *submitted personal data* is presented in-depth in [16]. Earlier, we highlighted the overarching finding that young and old males of lower educational degrees are particularly susceptible to phishing attacks. Additionally, we presented an overview of the underlying data in Table 1 for statistical insights into the different interactions, and Table 2 for the socio-demographic background of our study population. Insights into the qualitative data as obtained through analysis of survey answers and mapping to the HITL model for phishing investigation are presented in this manuscript. Using both data sets, we can quantify the success of a specific method of investigation in decreasing a participant's susceptibility to phishing attacks. We assessed correlations between all observed features throughout the three phases of the HITL model, *Notice*, *Expect & Suspect*, and *Investigate* and participants' reactions to our phishing emails, such as *link click* or *data submission*. The correlations were computed using primarily *chi-squared* tests ( $\chi^2$ )

<sup>3</sup> Once every 800ms, the user's browser sent all events that occurred in the past time-frame to our server. This included blur events of the webpage in case users placed the open tab in the background.



to identify significance in correlations and Spearman tests to identify the direction of impact for categorical variables. We use  $\alpha = 0.05$  as a quasi-standard for significance. Table 3 highlights and groups our analysis’s most essential and overarching observations.

**Table 3.** Highlights and overarching observations on elements noticed, or actions reported by participants throughout the HITL model, that correlate with interaction with our phishing emails. Full analysis available in Table 4 in the Appendix.

	Observation	Impact Sig.	$N_{true}$	Statistical Test		
				Type	Result	$p$
Notice	Noticing elements from email body (e.g., URLs, icons) correlates with increased interaction	$\wedge \checkmark$	103	$\chi^2$	15.909	0.000
	Noticing the senders’ name or address significantly correlates with decreased interaction	$\blacktriangledown \checkmark$	70	$\chi^2$	19.237	0.000
	Noticing the topic <i>parcel delivery</i> , significantly correlates with increased interaction	$\wedge \checkmark$	53	$\chi^2$	48.342	0.000
Expect & Suspect	Lack of personal context significantly correlates with decreased interaction	$\blacktriangledown \checkmark$	652	$\chi^2$	5.264	0.022
	Lack of knowledge of the entity significantly correlates with decreased interaction	$\blacktriangledown \checkmark$	35	$\chi^2$	9.798	0.002
	Lack of knowledge about sender significantly correlates with decreased interaction	$\blacktriangledown \checkmark$	16	$\chi^2$	5.941	0.015
	Not-expecting current shipments significantly correlates with decreased interaction	$\blacktriangledown \checkmark$	537	$\chi^2$	37.150	0.000
	Expecting shipments and deliveries significantly correlates with increased interaction	$\wedge \checkmark$	63	$\chi^2$	112.082	0.000
Investigate	Investigating <i>email headers</i> significantly correlates with decreased interaction	$\blacktriangledown \checkmark$	362	$\chi^2$	4.301	0.038
	User Y-axis scroll distance correlates with <b>not</b> submitting data on the webpage	$\blacktriangledown \checkmark$	1,549*	$t$	-9.3223	0.0000
	Webpage blur events ( <i>as casual proxy for user web searches</i> ) significantly correlates with <b>not</b> submitting any data	$\blacktriangledown \checkmark$	645	$\chi^2$	8.5307	0.0035

**Impact** refers to increased or decreased susceptibility of participants. Significance (**Sig.**) refers to whether the statistical evaluation reports significance given  $\alpha = 0.05$ .  
\* for the  $t$ -test,  $N$  refers to the entire amount of users that visited the webpage

## 7 Discussion and Contextualization of Results

The previous overview of the study results (Tab. 3) shows effects that require closer assessment and contextualization. In the following sections, we explore a few of the overarching measures applied and observations mentioned by participants during this study, which we observed to impact their susceptibility to phishing attacks.

## 7.1 Noticing, Expecting and Suspecting Context

*Contribution 2* We observe that identification of missing context during the phases *Expect* and *Suspect* significantly decreases participants' susceptibility to phishing attacks.

The different steps and phases of the HITL model are closely connected. Whenever a participant *notices* a specific feature, they automatically compare it to their *expectation*. If that differs, the participant *suspects* illegitimacy. Due to the posthoc nature of our survey, we expect that most participants only reported features that have caused particular suspicion (*Limitation 1*). Across the stages *Notice*, *Expect*, and *Suspect* of the HITL model, we generally observe that email features denoting context, such as the sender, the topic, and particularly the entity covered in the email impact participant's susceptibility to react to it. Participants observing that they had no connection to the entity generally performed better, as most of them interpreted the email as SPAM or unrelated to them and chose to ignore or delete the email.

Contrasting, participants for which the email fit into a current context, such as, e.g., they were currently expecting a delivery or recently ordered something on the Internet, usually performed worse, expecting a shipment significantly correlated with increased susceptibility ( $N = 63, \chi^2 = 112.082, p < 0.001$ ). Vice-versa, participants' awareness that they are not expecting a shipment significantly correlated with decreased susceptibility ( $N = 537, \chi^2 = 37.150, p < 0.001$ ).

## 7.2 Investigative Measures

*Contribution 3* Web Searches are one of the most successful investigation techniques in cases in which an email lacks technical indicators and thereby significantly decrease users' susceptibility to falling for phishing attacks.

Generally, upon *Suspicion* during looking at, e.g., an email, participants should start to investigate the nature of the email. Investigation is typically a process in which users attempt to gain more information about, e.g., the context of an email. In many teaching programs, measures such as the investigation of email headers, URL targets, and email bodies are named. We observed some participants who knew (and applied) these methods among our participants. However, we further observed investigation techniques such as *looking at the main webpage* or *clicking on the link*. Such methods can expose the participant to dangers upon visiting a malicious website.

Participants who investigated, e.g., email headers, were less likely to further interact with our phishing content. This observation is surprising, as the emails were not forged, and no manipulation that would have been visible in email headers was applied. Further, *Zheng et al.* reported that displaying email headers does not reduce phishing susceptibility as users often fail to interpret them correctly [38]. Instead, we judge that email header investigation is a casual proxy for participants to be more aware of potential cybersecurity risks, which was the reason for decreased interaction.

Considering further investigation techniques, participants who scrolled a lot on the webpage were significantly less likely to submit data on the webpage ( $N_{Visited} = 1,549, t = -9.3223, p < 0.001$ ). We interpret this measure as a proxy for users carefully interacting with and observing the webpage. In our questionnaire and free text answers specific to the emails, users reported investigation of the emails by web searches. Performing such web searches, users hoped to identify “*Who is actually behind the parcel service provider? Which corporation?*” (P365). Throughout the study, a total of 741 participants reported having performed web searches to retrieve more context on the supposed entities. Performing a Web search has proven to be one of the techniques significantly correlated with decreased submissions of private data on our webpage ( $N = 741, \chi^2 = 3.943, p = 0.047$ ). Overall, 13.73% of participants who did **not** perform web research submitted data in any of our phishing emails. In contrast, from those who performed web research, only 9.09% submitted personal data.

While we cannot track participant behavior after opening the email, we could track their behavior when visiting our webpage. One of the events we tracked is *webpage blur*, which occurs whenever a user selects any other tab in their browser but keeps our webpage open. We interpret webpage blur as a casual proxy for opening a new tab and performing a web search. This behavior significantly correlates with users not submitting data ( $N = 645, \chi^2 = 8.5307, p = 0.0035$ ).

### 7.3 Biases and Limitations

**(Non-) Technical Population Groups** One bias observed in our participant group (cf. Sec 5) is the difference between technical and non-technical people, e.g., expressed through jobs in IT and technology exposure. To test for the impact of IT affiliation, we compared different features throughout the HITL model between the two groups. We observe that the more technical features *noticed* by participants, such as the URL, are more often named among the IT population. In contrast, non-technical features, such as the personal salutation, were primarily noticed by people not affiliated with IT. Contextual aspects such as the email sender are noticed across both groups and contribute to not interacting with the email for both groups.

When assessing *Expected* and *Suspected* features inside the email, observations such as *Personal context significantly drives increased or decreased interaction* hold true independent of technicality. Similarly, throughout both groups, people who expected shipments, as reported in free text answers (cf. Figure 4), were more likely to click on the links and submit data on the web pages. One observed contrast between both groups is that IT-affiliated people expect and suspect more technical features, such as *parcel size*. However, that had no impact on increased or decreased susceptibility to phishing attacks in our study.

Regarding investigative measures, 352 IT-affiliated and 358 non-IT people have claimed to have performed (web-) searches. IT people were slightly better with their investigation, as only 8,89% of them submitted data, while 11,33% of non-IT people submitted data after having performed web searches. We observed advanced investigation techniques even among those participants who

were not affiliated with IT. However, among this group of people, we observed a higher error rate in interpreting the results (two of four non-IT participants have still submitted data after checking with, e.g., VirusTotal). We claim that this misjudgment stems from suboptimal training in which the absence of indicators for maliciousness (e.g., alerts in VirusTotal) is automatically interpreted as a positive sign without questioning if the measure applied (checking a URL with VirusTotal) is actually reasonable for the current assessment.

When *acting* on an email, after having investigated it, we assessed whether a participant clicked on the link or submitted data to the webpage. Further reactions, as reported during mapping participants' answers to the HITL model (cf. Sec. 6.1), included moving an email to SPAM or reporting it inside the company. We observed that reporting the email to any third party (IT department, police, other institutions) was more often reported among non-IT participants.

**Participant Group Recruitment** Our participants were recruited from the online education platform *openHPI*, on which free video-based online courses are offered. The platform mostly features educational courses on IT topics, such as programming, databases, AI, or cybersecurity. Earlier, we discussed observed differences between IT-affiliated participants and those participants who are employed, e.g., in jobs in public administration. Still, we assume that throughout all participants in all job roles, a particular interest in IT is apparent. Otherwise, they would not be enrolled in the platform. Therefore, we assume that the generalizability of our results towards the general public, particularly the non-IT population, is limited. Building on this thought, however, makes apparent that even those particularly interested in information technology (i.e., participants of our study) failed to correctly interpret indicators of mistrust such as missing context during our study. Replicating the study with a more representable population group would show less applied technical investigation techniques and an even higher failure rate to assess the study content as phishing correctly, matching results from related works.

**Debriefing of Participants** Our study targeted real-world participants who had not received the entire disclosure of the study context upon providing consent to participate (cf. Section 4.2). Hence, participants are likely unaware of participating in a phishing study. Therefore, we were required to ensure that our study emails would not be propagated further beyond our study participants. While the emails should withhold brief investigation by participants, any more technical investigation of our emails or web pages by trained staff should easily refer to the research context of the study. Therefore, we included debriefing information at various points throughout the study design:

1. The emails contained debriefing as white text on a white background
2. The imprint of the webpages contained debriefing information
3. Upon submitting data to the webpage, participants were debriefed
4. Upon contacting the sender of the emails, we debriefed the participants

The hidden debriefing information in emails potentially impacts users with screen readers or those who viewed the emails in plain text. For the former, investigating the behavior of users with disabilities would be part of a larger research question covering the impact of assistance tools on (phishing) assessment practices. An appropriate analysis of this research question is out of scope for our study. For the latter, in the survey, only a few people ( $< 10$ ) responded to have viewed the email in plain text. The practice is uncommon in our study population. As we can not track which participants have observed the debriefing information inside the email, we acknowledge the limitation. However, we must omit a detailed analysis of the practice in this manuscript.

With the different measures of debriefing in place, debriefed participants within each intervention needed to be excluded from the analysis. For example, for a user who submitted data and later further investigated the web page, that web page behavior should not be tracked and assessed. For debriefing measures two through four, we could technically track through requests to our web servers or emails to the contact addresses when the debriefing happened. In the data cleaning step three (Section 4.4), such participant behavior after debriefing was excluded from the dataset before further analysis.

## 8 Future Work

Our study targeted participants' investigative behavior of phishing content without traditional indicators for phishing. This increases the difficulty of email assessment, resulting in a relatively high time investment. To further study the human cost of phishing, a follow-up study using the same setup and observed data points could be developed studying phishing emails showing traditional cues for phishing, such as tampered email senders, illegitimate links, or spelling and grammar mistakes. Such analysis could provide interesting results, allowing further interpretation of the human cost of phishing attacks.

During our qualitative analysis of participants' answers, we observed the phenomenon of participants reporting, e.g., content in English, while all content was purely designed and distributed in German. This is likely because our survey was only sent posthoc, up to 3.5 months after the first phishing email (cf. study setup depicted in Figure 1). On the other hand, participants could also be subject to the phenomenon of confirmation bias and thus misremember information. Further research studying this observation could be performed by interlacing the *Staged Innovation Design* with surveys to some participants, after which those would be removed from the future participant pool to ensure maintaining an unbiased study population while retrieving intermediate survey answers.

## 9 Conclusion

This manuscript presents the results of a large-scale mixed-methods study examining human-phishing interaction when confronted with emails that lack traditional cues for phishing. We provide three *human-in-the-loop* model taxonomies

of 950 participants' phishing email investigation approaches. We observe the major contributor to phishing susceptibility in our study to be the identification of (missing) context. As expected, this is the only valid indicator for phishing in the study emails, as the fictitious entities had actual web pages, and no, e.g., links were manipulated. Participants unsure of the nature, entity, or subject of the emails reported to have performed web searches for further investigation. Verifying with our data on submissions of private data on the phishing web page, we could observe that the participants who mentioned having performed web research submitted sensitive data in 33.79% fewer cases than the cohort.

In our study, most users intuitively reacted well to the challenge of missing cues for phishing inside the emails. However, we also observed users who failed to make proper decisions. One reason might be users unaware of the implications of data disclosure to an attacker. We call on educators to highlight the risks of providing sensitive data to cybercriminals more prominently. Furthermore, concepts currently only employed in professional contexts, such as highlighting if an email is from an external organization, could also be beneficial in private contexts. E.g., a banner in email applications for emails where it is the first time the user has contact with the entity could help participants derive context.

Various qualitative answers by participants have shown that they assess their emails to know whether they are required to react. However, they need guidance and easily usable tools to support their investigation process. Hence, developing tools and measures to help laypersons investigate (phishing) emails securely should be prioritized in research and product development.

## References

1. Al-Daeef, M.M., Basir, N., Saudi, M.M.: Security awareness training: A review. *Lecture Notes in Engineering and Computer Science* (2017), iSBN: 2078-0958
2. Alharbi, A., Alotaibi, A., Alghofaili, L., Alsalamah, M., Alwasil, N., Elkhediri, S.: Security in Social-Media: Awareness of Phishing Attacks Techniques and Counter-measures. In: 2022 2nd International Conference on Computing and Information Technology (ICCIT) (2022). <https://doi.org/10.1109/ICCIT52419.2022.9711640>
3. Alzubaidi, A.: Measuring the level of cyber-security awareness for cybercrime in Saudi Arabia. *Heliyon* **7**(1) (2021). <https://doi.org/10.1016/j.heliyon.2021.e06016>
4. Caputo, D.D., Pfleeger, S.L., Freeman, J.D., Johnson, M.E.: Going Spear Phishing: Exploring Embedded Training and Awareness. *IEEE Security & Privacy* **12**(1), 28–38 (Jan 2014). <https://doi.org/10.1109/MSP.2013.106>
5. Cranor, L.F.: A framework for reasoning about the human in the loop (2008)
6. European Union Agency for Cybersecurity: ENISA Threat Landscape 2022 (2022), <https://www.enisa.europa.eu/publications/enisa-threat-landscape-2022>
7. Federal Bureau of Investigation: Business email compromise (2022), <https://www.fbi.gov/how-we-can-help-you/safety-resources/scams-and-safety/common-scams-and-crimes/business-email-compromise>
8. Fernando, M., Arachchilage, N.: Why Johnny can't rely on anti-phishing educational interventions to protect himself against contemporary phishing attacks? *ACIS 2019 Proceedings* (Jan 2019), <https://aisel.aisnet.org/acis2019/42>

9. Finn, P., Jakobsson, M.: Designing ethical phishing experiments. *IEEE Technology and Society Magazine* **26**(1), 46–58 (2007). <https://doi.org/10.1109/MTAS.2007.335565>, conference Name: IEEE Technology and Society Magazine
10. Furnell, S.: Phishing: can we spot the signs? *Computer Fraud & Security* **2007**(3), 10–15 (Mar 2007). [https://doi.org/10.1016/S1361-3723\(07\)70035-0](https://doi.org/10.1016/S1361-3723(07)70035-0)
11. Greitzer, F.L., Li, W., Laskey, K.B., Lee, J., Purl, J.: Experimental Investigation of Technical and Human Factors Related to Phishing Susceptibility. *ACM Transactions on Social Computing* **4**(2), 1–48 (Jun 2021). <https://doi.org/10.1145/3461672>
12. Innab, N., Al-Rashoud, H., Al-Mahawes, R., Al-Shehri, W.: Evaluation of the Effective Anti-Phishing Awareness and Training in Governmental and Private Organizations in Riyadh. In: 2018 21st Saudi Computer Society National Computer Conference (NCC). pp. 1–5 (Apr 2018). <https://doi.org/10.1109/NCG.2018.8593144>
13. Jampen, D., Gür, G., Sutter, T., Tellenbach, B.: Don’t click: towards an effective anti-phishing training. A comparative literature review. *Human-centric Computing and Information Sciences* **10**(1), 33 (Aug 2020). <https://doi.org/10.1186/s13673-020-00237-7>
14. Jensen, M.L., Dinger, M., Wright, R.T., Thatcher, J.B.: Training to Mitigate Phishing Attacks Using Mindfulness Techniques. *Journal of Management Information Systems* **34**(2), 597–626 (Apr 2017). <https://doi.org/10.1080/07421222.2017.1334499>, publisher: Routledge
15. Kumaraguru, P., Cranshaw, J., Acquisti, A., Cranor, L., Hong, J., Blair, M.A., Pham, T.: School of phish: a real-world evaluation of anti-phishing training. In: *Proceedings of the 5th Symposium on Usable Privacy and Security*. pp. 1–12 (2009)
16. Köhler, D., Pünter, W., Meinel, C.: Fishing for non-professional answers: Quantitative study on email phishing susceptibility in private contexts (2023). <https://doi.org/10.13140/RG.2.2.21865.47201/1>, in Review
17. Meinel, C., Willems, C., Staubitz, T., Sauer, D., Hagedorn, C.: openHPI: 10 Years of MOOCs at the Hasso Plattner Institute (2022)
18. Mitnick, K.D., Simon, W.L.: *The art of deception: Controlling the human element of security*. John Wiley & Sons (2003)
19. Nguyen, C., Jensen, M., Day, E.: Learning not to take the bait: a longitudinal examination of digital training methods and overlearning on phishing susceptibility. *European Journal of Information Systems* **32**(2), 238–262 (Mar 2023). <https://doi.org/10.1080/0960085X.2021.1931494>
20. Nthala, N., Wash, R.: How Non-Experts Try to Detect Phishing Scam Emails. *Workshop on Consumer Protection* (May 2021), <https://par.nsf.gov/biblio/10297019-how-non-experts-try-detect-phishing-scam-emails>
21. Parsons, K., Butavicius, M., Pattinson, M., McCormac, A., Calic, D., Jerram, C.: Do Users Focus on the Correct Cues to Differentiate Between Phishing and Genuine Emails? *ACIS 2015 Proceedings* (Jan 2015), <https://aisel.aisnet.org/acis2015/6>
22. Parsons, K., McCormac, A., Pattinson, M., Butavicius, M., Jerram, C.: Phishing for the Truth: A Scenario-Based Experiment of Users’ Behavioural Response to Emails. In: *Security and Privacy Protection in Information Processing Systems*. pp. 366–378. *IFIP Advances in Information and Communication Technology*, Springer, Berlin, Heidelberg (2013). [https://doi.org/10.1007/978-3-642-39218-4\\_27](https://doi.org/10.1007/978-3-642-39218-4_27)
23. Rajivan, P., Gonzalez, C.: Creative Persuasion: A Study on Adversarial Behaviors and Strategies in Phishing Attacks. *Frontiers in Psychology* **9** (2018)
24. Resnik, D.B., Finn, P.R.: Ethics and Phishing Experiments. *Science and Engineering Ethics* **24**(4), 1241–1252 (Aug 2018). <https://doi.org/10.1007/s11948-017-9952-9>

25. Schroeder, J.: Advanced Persistent Training: Take Your Security Awareness Program to the Next Level. Apress (Jun 2017), google-Books-ID: UjgoDwAAQBAJ
26. Siadati, H., Palka, S., Siegel, A., McCoy, D.: Measuring the Effectiveness of Embedded Phishing Exercises (2017), <https://www.usenix.org/conference/cset17/workshop-program/presentation/siadatii>
27. Stockhardt, S., Reinheimer, B., Volkamer, M., Mayer, P., Kunz, A., Rack, P., Lehmann, D.: Teaching Phishing-Security: Which Way is Best? In: ICT Systems Security and Privacy Protection. pp. 135–149. IFIP Advances in Information and Communication Technology, Springer International Publishing, Cham (2016). [https://doi.org/10.1007/978-3-319-33630-5\\_10](https://doi.org/10.1007/978-3-319-33630-5_10)
28. Sutter, T., Bozkir, A.S., Gehring, B., Berlich, P.: Avoiding the Hook: Influential Factors of Phishing Awareness Training on Click-Rates and a Data-Driven Approach to Predict Email Difficulty Perception. IEEE Access **10**, 100540–100565 (2022). <https://doi.org/10.1109/ACCESS.2022.3207272>
29. The MITRE Corporation: CAPEC-98: Phishing (2021), <https://capec.mitre.org/data/definitions/98.html>
30. UNESCO Institute for Statistics: International standard classification of education: Isced 2011 (2012), <https://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf>
31. United Nations Department of Economic and Social Affairs: International standard industrial classification of all economic activities (2008), [https://unstats.un.org/unsd/publication/SeriesM/seriesm\\_4rev4e.pdf](https://unstats.un.org/unsd/publication/SeriesM/seriesm_4rev4e.pdf)
32. Wagner, N.: Instructional product evaluation using the Staged Innovation Design. Journal of Instructional Development **7** (1984)
33. Wash, R.: How Experts Detect Phishing Scam Emails. Proceedings of the ACM on Human-Computer Interaction **4** (2020). <https://doi.org/10.1145/3415231>
34. Wash, R., Cooper, M.M.: Who Provides Phishing Training? Facts, Stories, and People Like Me. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, New York (2018). <https://doi.org/10.1145/3173574.3174066>
35. Wash, R., Nthala, N., Rader, E.: Knowledge and Capabilities that {Non-Expert} Users Bring to Phishing Detection. pp. 377–396 (2021), <https://www.usenix.org/conference/soups2021/presentation/wash>
36. Wen, Z.A., Lin, Z., Chen, R., Andersen, E.: 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. pp. 1–12. CHI '19, ACM, New York, USA (May 2019). <https://doi.org/10.1145/3290605.3300338>
37. Williams, E.J., Polage, D.: How persuasive is phishing email? The role of authentic design, influence and current events in email judgements. Behaviour & Information Technology **38** (Feb 2019). <https://doi.org/10.1080/0144929X.2018.1519599>
38. Zheng, S., Becker, I.: Presenting suspicious details in user-facing e-mail headers does not improve phishing detection. In: SOUPS @ USENIX Security Symposium (2022), <https://api.semanticscholar.org/CorpusID:252996739>
39. Zheng, S.Y., Becker, I.: Checking, nudging or scoring? Evaluating e-mail user security tools. pp. 57–76 (2023), <https://www.usenix.org/conference/soups2023/presentation/zheng>



## A Appendix: Survey Instrument

*The survey questions are translated from German for publication in this manuscript. The following sub-sections layout the survey instrument used to obtain the responses presented throughout the manuscript.*

### A.1 Demography

- Q1: Please enter your email address.  
 Q2: How old are you?  
 Q3: Which gender would you associate yourself with?  
 Q4: Which is your highest level of education?  
 Q5: In which industry are you currently working? (*Multi-Select among primary industry groups according to UN ISIC Rev.4 [31]*)

### A.2 Phishing Emails and Reactions

- Q6: In the past 4 months, we have sent 4 phishing emails as part of this study. In the following questions, we would like to know whether and how you reacted to the corresponding emails. You can view the four emails again here:  
 Q7: Have we successfully persuaded you to enter data during our campaign?  
 Q8: Which of the four phishing emails do you remember? (*Multi-Select*)  
 Q9: What was the major reason for a reaction to the email? (*Matrix-Select, one reason per email*)
- Curiosity
  - Fear
  - Pressure
  - Financial Interest
  - Trust
  - Authority
  - *I did not react to this mail.*
  - *Prefer not to answer.*
- Q10: Please provide more information on your reaction. (*Freertext*)  
 Q11: Which of the emails gave you the feeling that something was wrong? (*Multi-Select*)  
 Q12: Please explain your feelings on the emails. (*Freertext answer for each email*)  
 Q13: What did you do when you had off feelings with an email? (*Multi-Select*)
- Visit the main webpage
  - Perform a web search
  - View the website imprint
  - View the website data privacy declaration
  - Investigate the website source code
  - Investigate the link target
  - Investigate the sender
  - Investigate the email header
- Q14: Have you carried out any further checks? (*Freertext*)  
 Q15: Which precautions have you taken for your investigation? (*Multi-Select*)
- I did not take special precautions.
  - VPN

- TOR
- Deactivate JavaScript
- Deactivate Cookies
- Use a special browser
- Use a sandbox / virtual machine
- Issue WHOIS / RDAP request for the IP / domain

**Q16:** Did you implement any other precautions or technical measures? (*Freertext*)

### A.3 IT-Context and Sensitization

**Q17:** How often do you use IT-Devices for your work and in your leisure time?

**Q18:** Estimate, how many emails you receive per day in your private and work contexts.

**Q19:** Did you previously participate in courses or training for cybersecurity awareness?

**Q20:** Which types of trainings did you previously participate in? (*Multi-Select*)

- Classroom training (including digital group training)
- Awareness information emails
- Test phishing emails (outside this study)
- Computer-based training
- Online courses
- Information videos
- Social media content
- Documentations (TV, Youtube)
- Podcasts and radio
- Print media (newspapers, flyer)
- Posters and billboard advertisement
- Other (*Freertext*)

**Q21:** How long ago did you participate in your last training?

**Q22:** Have you previously been affected by a security incident? (*Multi-Select*)

- Reacted to a phishing email
- Malware infection
- Lost a password
- Lost data
- Lost access to an account
- Stolen devices
- Lost money
- Other (*Freertext*)

## B Appendix: Large Scale Images of Phishing Content

The paper incorporates tiny graphics as an overview of the emails and webpages employed throughout the four iterations of our phishing study. Here, we provide the following images for readers who want to look at larger-scale variants.

## C Appendix: HITL-Model: Figures

Presented in the paper were shortened versions of the two taxonomies that highlight aspects which were more frequently named by study participants. However, in case fellow researchers would be designing similar studies, even answers from single participants could be helpful to understand what behavior to expect. Therefore, we present Figures 9 and 10, showing the full range of participant responses to the survey on the respective stages in the HITL model.

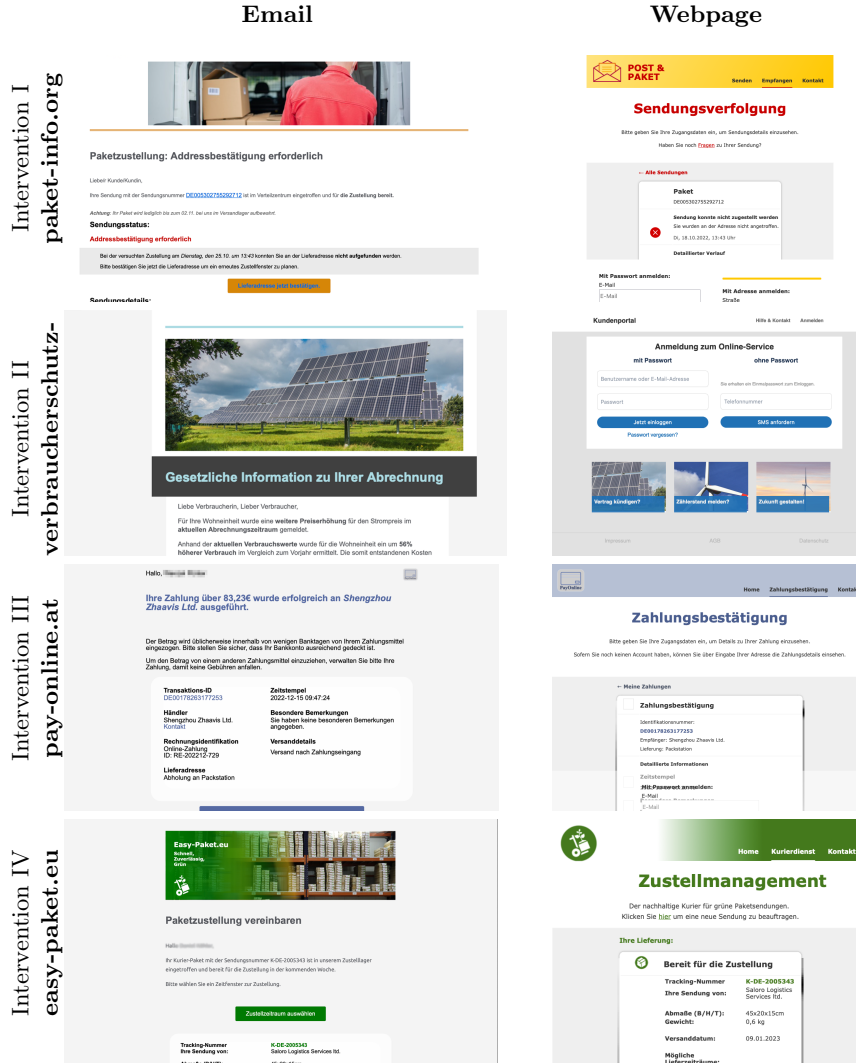
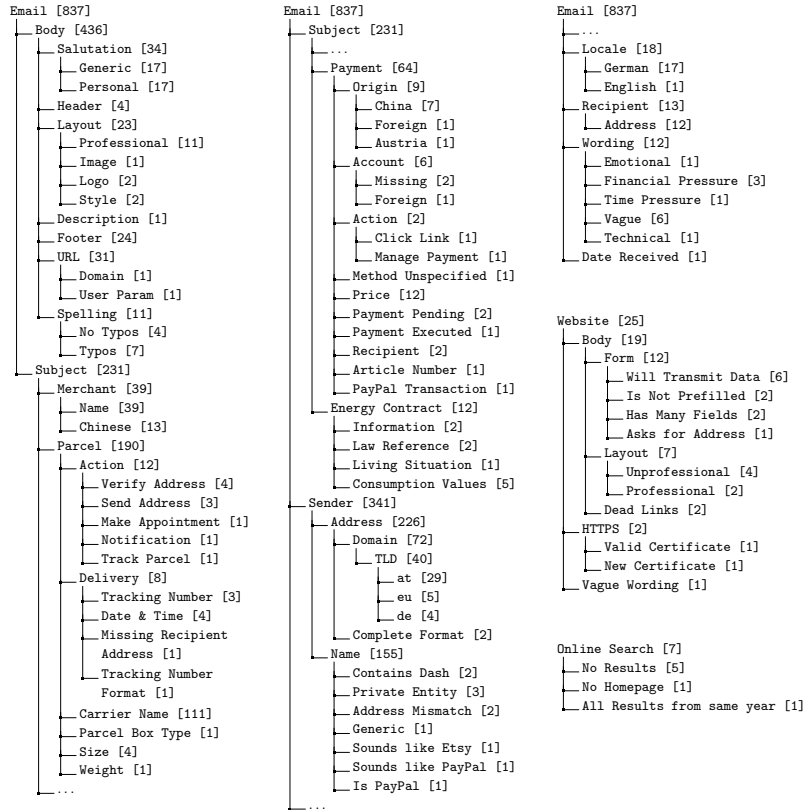


Fig. 8. Large-scale screenshots of German phishing content sent throughout the four iterations.

## D Appendix: Resulting Correlations

In Table 3, we summarized the most important correlations we observed between our participant responses and their interaction with our phishing emails and web pages. The analysis has brought us to identify the highlighted observations as particularly important, e.g. because we further observed mentions of the aspects in qualitative answers. Additionally, Table 4 provides an overview of all impactful aspects derived during our analysis. In the table, we group the findings by

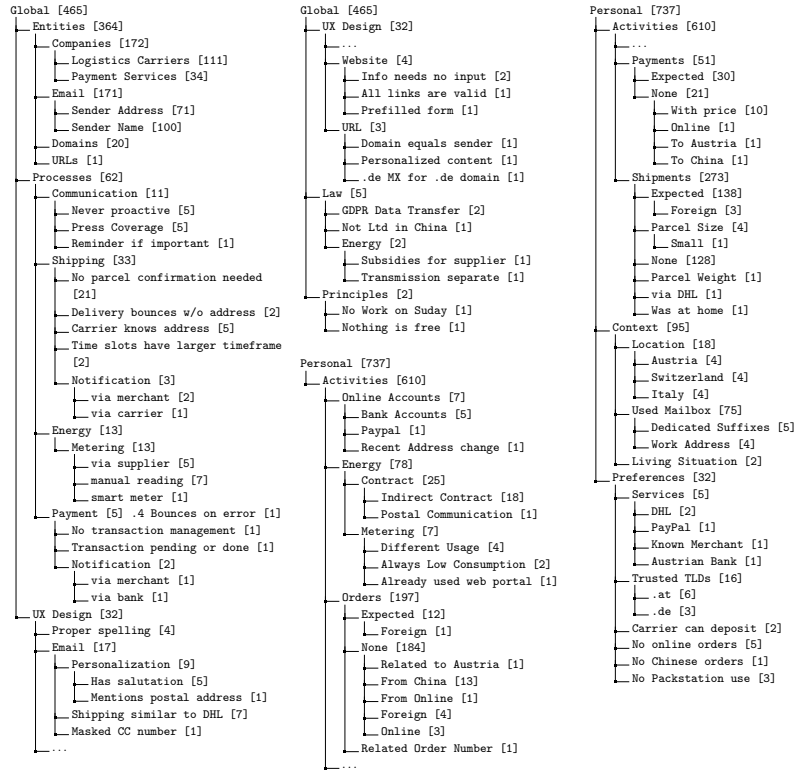


**Fig. 9.** Hierarchical overview of *noticed* properties named in survey responses. Participants *noticed* aspects within the *Email*, *Website* and during their *Online Searches*.

\* Numbers in [brackets] refer to the count of mentions.

the iteration they were reported of, with *General* applying to answers given to general, overarching question not directly targeted towards single interventions. Inside each iteration, we differentiate between the different phases of the HITL interaction model: *Notice* (N), *Expect* (E), *Suspect* (S), *Investigate* (Inv.), and *Act* (A). We compare the performance of the group that reported the respective feature (*Share of Participants*) to the performance of the *General Population*. Depending on whether participants who reported the respective feature performed better or worse, we indicate whether the respective group of participants *reacted* (React.) more or less often than their peers. The reaction translates to the phishing susceptibility, as indicated in Table 3 in the manuscript. An increased amount of reaction and, thereby, increased susceptibility hereby indicates worse behavior. Below, we provide one example of how to read the table:

*Reading Example:* People that highlighted *General Expectations* towards how (third party) entities perform email communication (global/entities/email)



**Fig. 10.** Hierarchical overview of *expected* and *suspected* properties named in survey responses. We differentiate between *Global* expectations that could be identical across participants and *Personal* expectations, such as a concrete shipment.

\* Numbers in [brackets] refer to the count of mentions.

performed better than their peers. Out of the 32 people who highlighted the respective feature, only 3.1% *Clicked* on the links provided in the emails, while generally, 24% of participants clicked on the links provided. This observation is statistically significant, as confirmed with a  $\chi^2$  test for significance resulting in  $p = 0.001$ .

**Table 4.** Overview of all Correlations observed between Participant Responses clustered to the HITL model.

HITL	HITL Feature	Share of		General Population	React. Sig.	$\chi^2$ Test		
		Participants	Interaction			$N_{true}$	Result	$p$
Notice (N)	email/sender/name	12.5%	Clicked	24.0%	▼ ✓	24	6.283	0.012
	email/sender	17.1%	Clicked	24.0%	▼ ✓	70	19.237	0.000
	email/sender	8.6%	Submitted	7.0%	^ ✓	70	9.268	0.002
	email/body	50.5%	Clicked	24.0%	^ ✓	103	15.909	0.000
	email/body	32.0%	Submitted	7.0%	^ ✓	103	17.211	0.000
	email/body/parcel	77.4%	Clicked	24.0%	^ ✓	53	48.342	0.000
	email/body/parcel	52.8%	Submitted	7.0%	^ ✓	53	44.334	0.000
	email/body/footer	11.5%	Clicked	24.0%	▼ ✓	26	7.583	0.006
	email/sender/address	17.3%	Clicked	24.0%	▼ ✓	52	11.744	0.001
	email/sender/address	7.7%	Submitted	7.0%	^ ✓	52	6.567	0.010
General & Overall Expect (E)	personal/activities	31.6%	Clicked	24.0%	^ ✓	636	4.925	0.026
	personal/activities	13.1%	Submitted	7.0%	^ ✓	636	6.390	0.011
	personal	31.4%	Clicked	24.0%	^ ✓	652	5.264	0.022
	personal	12.7%	Submitted	7.0%	^ ✓	652	3.891	0.049
	personal/activities/shipments/none	24.6%	Clicked	24.0%	^ ✓	537	37.150	0.000
	personal/activities/shipments/none	6.7%	Submitted	7.0%	▼ ✓	537	63.413	0.000
	global/entities/email/sender	0.0%	Clicked	24.0%	▼ ✓	16	5.941	0.015
	global/processes/energy/metering	100.0%	Clicked	24.0%	^ ✓	3	3.866	0.049
	global/processes/energy	100.0%	Clicked	24.0%	^ ✓	3	3.866	0.049
	global	0.0%	Submitted	7.0%	▼ ✓	44	5.436	0.020
	global/entities/email	3.1%	Clicked	24.0%	▼ ✓	32	10.850	0.001
	global/entities	5.7%	Clicked	24.0%	▼ ✓	35	9.798	0.002
	global/entities	0.0%	Submitted	7.0%	▼ ✓	35	4.043	0.044
	personal/activities/shipments/expected	90.5%	Clicked	24.0%	^ ✓	63	112.082	0.000
	personal/activities/shipments/expected	66.7%	Submitted	7.0%	^ ✓	63	184.949	0.000
	legit/parcel/other-order	100.0%	Clicked	24.0%	^ ✓	6	4.762	0.029
	legit/parcel/other-order	100.0%	Submitted	7.0%	^ ✓	6	8.584	0.003
	legit/parcel	76.9%	Clicked	24.0%	^ ✓	26	14.182	0.000
	legit/parcel	65.4%	Submitted	7.0%	^ ✓	26	16.313	0.000
Suspect (S)	legit	65.6%	Clicked	24.0%	^ ✓	32	7.594	0.006
	legit	53.1%	Submitted	7.0%	^ ✓	32	8.100	0.004
	fraud	23.5%	Clicked	24.0%	▼ ✓	17	5.829	0.016
	fraud	11.8%	Submitted	7.0%	^ ✓	17	5.835	0.016
	legit/payment	0.0%	Submitted	7.0%	▼ ✓	8	4.000	0.046
	fraud/identity	0.0%	Submitted	7.0%	▼ ✓	8	4.000	0.046
	Inv. info/search	29.3%	Clicked	24.0%	^ ✓	741	4.191	0.041
	Inv. info/website/link	64.7%	Clicked	24.0%	^ ✓	17	7.550	0.006
	Act react	64.3%	Clicked	24.0%	^ ✓	14	9.820	0.002
	Act react	42.9%	Submitted	7.0%	^ ✓	14	20.560	0.000
Iteration 2 Expect	email/sender/address/domain/tld	50.0%	Submitted	2.9%	^ ✓	2	5.508	0.019
	email/sender/name	3.4%	Clicked	18.2%	▼ ✓	29	4.050	0.044
	email/sender/address/domain/tld/de	50.0%	Submitted	2.9%	^ ✓	2	5.508	0.019
	personal/preferences/trusted-tld	50.0%	Submitted	2.9%	^ ✓	2	5.512	0.019
	personal/preferences/trusted-tld/at	50.0%	Submitted	2.9%	^ ✓	2	5.512	0.019
	personal/preferences	50.0%	Submitted	2.9%	^ ✓	2	5.512	0.019
	Inv. verify/energy-law	66.7%	Clicked	18.2%	^ ✓	3	10.616	0.001
Iteration 3 Expect	verify/energy-law	33.3%	Submitted	2.9%	^ ✓	3	5.876	0.015
	info/search/company	100.0%	Submitted	2.9%	^ ✓	1	19.502	0.000
	Notice website	33.3%	Submitted	2.8%	^ ✓	3	4.742	0.029
	Notice website/wording	100.0%	Submitted	2.8%	^ ✓	1	16.000	0.000
	Notice website/wording/vague	100.0%	Submitted	2.8%	^ ✓	1	16.000	0.000
	Inv. info/search	7.9%	Clicked	19.5%	▼ ✓	547	5.923	0.015
	Notice website	62.5%	Clicked	31.1%	^ ✓	8	6.593	0.010
Iteration 4 Expect	website/body	62.5%	Clicked	31.1%	^ ✓	8	6.593	0.010
	website/body/form	66.7%	Clicked	31.1%	^ ✓	6	5.483	0.019
	global/entities	6.6%	Submitted	12.7%	▼ ✓	137	4.597	0.032
	global	6.2%	Submitted	12.7%	▼ ✓	144	3.920	0.048
	global/entities/email/sender	13.6%	Submitted	12.7%	^ ✓	22	3.982	0.046
	personal	1.5%	Submitted	12.7%	▼ ✓	199	6.091	0.014
	personal/activities/shipments/expected	45.8%	Clicked	31.1%	^ ✓	24	12.202	0.000
	Inv. info/search	15.1%	Clicked	31.1%	▼ ✓	535	18.303	0.000
	info/website/imprint	45.0%	Clicked	31.1%	^ ✓	20	9.208	0.002
	info/website	48.1%	Clicked	31.1%	^ ✓	27	16.680	0.000

*Reaction (React.)* refers to whether the respective group reacted more or less than their peers.  
*Significance (Sig.)* refers to whether the statistical evaluation reports significance given  $\alpha = 0.05$ .