"Do Users fall for Real Adversarial Phishing?" Investigating the Human response to Evasive Webpages

Ajka Draganovic*, Savino Dambra¶, Javier Aldana Iuit§, Kevin Roundy¶, Giovanni Apruzzese*
*University of Liechtenstein, ¶Norton Research Group, §Avast Software
{name.surname}@{uni.li*, gendigital.com§¶},

***Abstract*—Phishing websites are everywhere, and countermeasures based on static blocklists cannot cope with such a threat. To address this problem, state-of-the-art solutions entail the application of machine learning (ML) to detect phishing websites by checking if they visually resemble webpages of well-known brands. These techniques have achieved promising results in research and, consequently, some security companies began to deploy them also in their phishing detection systems (PDS). However, ML methods are not perfect and some samples are bound to bypass even production-grade PDS.**

In this paper, we scrutinize whether *genuine phishing websites* that evade *commercial ML-based PDS* represent a problem "in reality". Although nobody likes landing on a phishing webpage, a false negative may not lead to serious consequences if the users (i.e., the actual target of phishing) can recognize that "something is phishy". Practically, we carry out the first user-study (N=126) wherein we assess whether unsuspecting users (having diverse backgrounds) are deceived by "adversarial" phishing webpages that evaded a real PDS. We found that some well-crafted adversarial webpages can trick most participants (even IT experts), albeit others are easily recognized by most users. Our study is relevant for practitioners, since it allows prioritizing phishing webpages that simultaneously fool (i) machines and (ii) humans—i.e., their intended targets.

I. Introduction

The battle against phishing is still ongoing [1], despite decades of efforts aimed at countering this threat [2, 3]. According to the FBI, phishing is the leading form of cyber-crime [4], and its proliferation is constantly increasing [5].

Phishing *websites* are among the most common vectors used by adversaries to carry out phishing attacks [6]. After deploying their phishing hooks "in the wild," attackers try to lure their victims (through, e.g., social engineering) to such malicious webpages—intent to steal their private data, or compromise their IT systems. Countermeasures to phishing can fall in two categories: *human-centered* (e.g., phishing awareness training [7]), which aim at improving the ability of humans to avoid phishing traps; and *machine-centered* (e.g., phishing website detectors [8]), which aim at preventing the human user from landing on a phishing trap in the first place. As a matter of fact, the fight against phishing can be seen as a **two-step decision process**, which we illustrate in Fig. 1. After a user is brought to any given website, a phishing detection system (PDS) quickly analyzes the website (e.g., by checking some blocklists or using heuristics [9]): if the PDS determines the website to be phishing, then the webpage is not displayed to the user (who might be shown a warning/alert); otherwise, the browser renders the webpage. Of course, no issue arises if the webpage is benign. However, if the webpage is malicious, the decision is now up to the user: if they can recognize the page as phishing, then the attack is defused; otherwise, the user (i.e., its data or device) may be "caught".

Unfortunately, *operational* PDS are tweaked to minimize the rate of false alarms, which leads to a significant number of phishing websites to *evade* their detection (a security company had over 9k "evasions" in just one month [10]).[1] Given the brittleness of existing anti-phishing schemes, it is paramount to improve the users' ability to autonomously recognize phishing websites. However, according to a recent Proofpoint's report [1], more than 33% companies do not have any training program; and, among those which do provide such training, more than 50% do so via simulations.

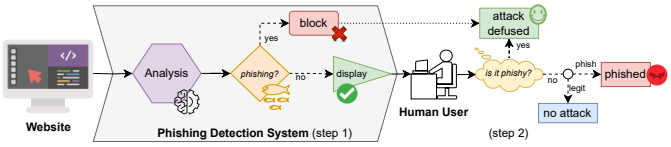


Fig. 1: Scenario: phishing detection is a two-step decision process.

Amidst this chaos, we observe that there is a **mismatch between research efforts that focus on human- or machine-centered** solutions. In particular, despite phishing detection being a two-step decision process, prior work only focused on either one of these steps. For example, papers that propose novel PDS tend to overlook how humans respond to those webpages that bypassed the proposed PDS; whereas papers that focus on the human perception of phishing websites do not consider webpages that evaded operational PDS (we discuss related work in §II-C). Such a disconnection is problematic: a PDS whose false negatives can trick all end users (i.e., the true target of phishing) is not reliable; whereas carrying out phishing assessments using webpages that can be trivially blocked by PDS may not be the best way to invest resources.

In this paper, we seek to bridge the gap between these two complementary approaches against phishing. To this end,

[1]This is why PDS have been relying on blocklists for a long time [11], albeit state-of-the-art PDS now also leverage artificial intelligence to provide an additional layer of defense [10, 12]. See §II-A for background.

we reach out to a security company that develops anti-phishing schemes, and obtain a set of "adversarial phishing webpages" (AW) that evaded their operational PDS powered by deep learning (§III-A). Then, we carry out a user-study (N=126) in which we ask participants (who were not primed in any way) to figure out whether such AW resemble legitimate websites or not (§III-B). We also inquire potential explanations for their skepticism (if any). We analyze our results both quantitatively (§IV) and qualitatively (§VI). Our findings reveal that while "poorly crafted" AW can be easily spotted by end-users, others can deceive (almost) all of our participants.

CONTRIBUTION. To bridge the gap between human- and machine-centered anti-phishing schemes, we:

- Carry out the *first* user-study elucidating the response of humans to *real* phishing webpages that evaded a *real* phishing detection system based on deep learning.
- Validate our findings quantitatively—via *statistical* tests (§V-A); and qualitatively—alongside *practitioners* (§VI-C); and derive *recommendations* for research (§VII-B).
- Provide *practical insights* on operational PDS (§III-A) and on how to improve them (we share our phishing data [13]);

This study can spearhead future work aimed at improving PDS, i.e., by identifying the AW that deceive most users, and then fixing PDS so that such AW are not misclassified.

II. Background and Motivation

We focus on phishing *websites*. Other forms of phishing (e.g., email [14–16]) are outside our scope—albeit our findings can apply to these (if they envision luring a victim to a website).

To allow a complete understanding of the problem tackled by our paper (§II-B), we summarize the landscape of phishing website detection (§II-A), and then position our paper within existing literature on the human perception of phishing (§II-C).

A. Phishing Website Detection

The first line of defense against phishing entails *automated* detection schemes [17]. The goal is analysing a given website to determine whether it is malicious (or not) and, if so, prevent its webpage from being displayed to an end-user.

Rule-driven detection. The most popular way to fight phishing websites is through *blocklists* [11, 18, 19]: by checking if an URL (or part of it) is included in a pre-defined set of malicious URLs (or domains), it is possible to precisely (and quickly) identify phishing websites. Detection mechanisms based on blocklists are widespread in modern browsers (e.g., Google Safe Browsing [9]), and are appreciated due to their near-zero false positive rate. Unfortunately, even though such mechanisms are kept up-to-date with new malicious entries, these tools are useless against "novel" phishing websites [20].

Data-driven detection. To protect users against phishing websites that are not included in any blocklist, abundant research efforts proposed data-driven solutions based on heuristics. For instance, Zhang et al. [21] identified some patterns commonly associated to phishing, and used these to discriminate benign from phishing websites. Similar detection techniques also encompass machine learning (ML) methods [22]. For instance, Mohammad et al. [23] proposed a set of features (extracted from the URL and the HTML of a webpage) that could be used to develop a ML-based detector (after undergoing a proper training phase), whereas Cui et al. [24] use ML to infer malicious domains typically associated to phishing. A complementary data-driven approach against phishing entails using the *visual similarity*. Early works date back to 2006 [25], and some even leverage ML (e.g., [17]). More recently, due to the never-ending advancement of deep learning (DL), detection methods reliant on visual cues attracted much attention in research [8, 26–28]. However, despite abundant scientific literature, these proposals suffer from a significant drawback: the "high" false positive rate—which impairs the end-user experience. Consequently, the integration of ML/DL into operational phishing detection systems (PDS) proceeds at a slow pace—but it is happening [12].

Adversarial phishing. Besides having to deal with false positives, *real* PDS must face another issue: the "false negatives" that stem from adversaries who deliberately want to evade the PDS [3]. Indeed, abundant evidence suggests that even data-driven solutions cannot "catch-all-phish". Liang et al. [29] cracked Google's page filter in 2016. More recently, security enthusiasts bypassed the ML-based detector of a popular anti-phishing evasion competition [30]; Apruzzese et al. [31] showed that state-of-the-art detectors (analyzing either the URL/HTML) can be fooled via cheap perturbations, whereas Lee et al. [32] evaded logo-based detectors proposed in research with imperceptible visual changes. Finally, a recent work [10] showed that even *commercial-grade* PDS that uses DL for visual similarity exhibits thousands of false negatives every month—some of which due to "perturbations" that are easily recognizable by humans.

B. Problem Statement (Focus of the Paper)

Detecting phishing websites is tough, and false positives/negatives are bound to occur. In reality, however, practitioners are vexed by a dilemma: "what to prioritize?" [10]. The answer should be driven by the **perspective of the end-user**—the true target of phishing.

Nobody wants their browsing activities to be frequently interrupted by inaccurate blocking mechanisms (i.e., false positives). However, by turning the attention to the *false negatives* of a PDS (which lead to displaying a malicious page), there are two cases:

1) the user *recognizes* the page as phishing. Despite being an annoyance ("yet another phishing website!"), the consequences of such a misclassification are mild—the user will simply close the webpage and resume their activities.
2) the user *does not* recognize the page as phishing. This is a serious problem, since it may lead to the phishing attack being successful—causing a much greater loss (in terms of time, finance, or privacy [33]) to the user.

In this paper, we are inspired by such "dual nature" of **false negatives** in the context of phishing website detection. We seek to scrutinize the response of humans to those "adversarial webpages" that evaded a ML/DL-based PDS. From a practical

viewpoint, the findings of our study can assist practitioners in optimizing their (limited) resources, e.g., by placing more emphasis on those pages that can deceive users. To the best of our knowledge, we are the first to investigate this problem.

C. Related Work (User Studies)

Let us discuss our study in light of existing literature.

Technical papers. Abundant research on phishing website detection entail "technical papers", typically proposing either a defensive solution that improves the state-of-the-art (e.g., [34, 35]); or a new attack that bypasses existing anti-phishing schemes (e.g., [31]). Unfortunately, most such papers overlook the user perspective; with two notable exceptions: (1) Abdelnabi et al. [8], after proposing a novel detection technique, carried out a user-study asking participants to "evaluate how trustworthy [misclassified webpages] seem based on visual similarity". From a realistic viewpoint, however, such a user-study has two limitations: (i) it is based on the output of a research proposal—and not on a real PDS; and (ii) the phrasing of the question primes the users—who are more likely to be suspicious of a webpage, and which inevitably leads to biased results. (2) Lee et al. [32] propose a new means to evade PDS based on logo-identification, and then carry out a user-study wherein humans are asked to identify how similar "adversarial logos" are to the original logos of well-known brands. However, besides being also based on a research proposal, this user-study only focuses on the logo—which is only a small part of a real webpage, and is hence inappropriate to derive whether the user would be truly fooled by the corresponding webpage.

Human-centered papers. Differently from technical papers, another branch of research specifically focuses on investigating the human factor in phishing (for, e.g., educational training campaigns [6]). Such "human-centered" papers are closer to our work. However, none of these papers investigate (neither explicitly nor implicitly) the specific problem tackled in our research. To position our paper within related works, we carry out an extensive literature review[2] wherein we scrutinize each related work according to four criteria:

- **Deployed ML misclassifications:** did the phishing webpages in the user-study evade a real ML-based detector?
- **No priming:** were the participants kept in the dark about the study being about phishing? (otherwise, it can bias results)
- **Real phishing:** were the phishing webpages taken from "the wild web"? (perhaps such pages were created in a lab)
- **IT expertise:** was the IT expertise accounted for? (experts in IT may respond differently than amateurs)

These criteria allow one to assess the "realistic value" of the findings of each prior user-study. We summarize our literature review in Table I, in which we also report the amount of participants included in the user-study.

[2] We perform the literature review between November 2022 and June 2023. During this time-frame, two authors manually queried popular scientific repositories for user-studies on the perception of humans to phishing websites. The authors frequently met and discussed their individual findings across various meetings to derive a consensus. We omitted three works ([7, 36, 37]) because they focus on websites and emails—which are outside our scope.

TABLE I: User-studies on the human perception of phishing websites. A "?" denotes works for which we could not find any information.

Paper	Year	Sample size	Deployed ML missclass.	No Priming	Real Phishing	IT Expertise
Dhamija [38]	2006	22	✗	✗	✗	✓
Tsow [39]	2007	398	✗	✗	✗	?
Sheng [40]	2007	42	✗	✗	✓	✓
Jakobsson [41]	2007	400	✗	✗	?	✗
Herzberg [42]	2008	23	✗	✗	✗	✗
Alnajim [43]	2009	36	✗	✓	✓	✓
Kumaraguru [44]	2010	28	✗	✗	✓	✓
Yang [45]	2012	62	✗	✗	✗	✗
Asanka [46]	2013	40	✗	✗	✗	✓
Purkait [47]	2014	621	✗	✗	✗	✓
Scott [48]	2014	66	✗	✗	✗	✗
Alsharnouby [49]	2015	21	✗	?	✗	✓
Kunz [50]	2016	32	✗	✗	?	✗
Arachilage [51]	2016	20	✗	✗	✓	✓
Xiong [52]	2017	320	✗	✓	✗	✓
Moreno [53]	2017	175	✗	✗	✗	✗
Gopavaram [54]	2021	250	✗	✗	✗	✓
This Paper	2023	126	✓	✓	✓	✓

Considerations. By observing Table I, we can see that most papers do not meet all such criteria. In particular, no paper considered "Deployed ML misclassifications": this is not surprising, given that most PDS are closed-source (hence it is not known whether they integrate some ML or not) and deployment of ML in cyber security proceeds at a slow-pace [22]. We also mention that many studies tend to prime their participants, which may not represent a realistic scenario (if a user "expects" to encounter phishing, they are less likely to fall for it in the first place [55]).

> **Summary:** prior user-studies do not allow investigating the real response of humans to real phishing webpages that evaded a real ML-based phishing detection system.

Why ML? We are aware that many data-driven methods exist to fight phishing (§II-A). We focus on ML-based PDS (using visual similarity) due to their emerging deployment in the real-world [10], in an attempt to raise the awareness on the concrete issues of these schemes before they become widespread—given how easily they can be bypassed [10, 32].

III. Research Method

Our study revolves around a central research question (RQ): "How do users respond to phishing webpages that evaded a ML-based PDS"? To answer our RQ, we first obtain a set of adversarial phishing webpages (§III-A), and then devise a questionnaire (§III-B) aimed at assessing the awareness of users (§III-C) to such webpages.

A. Data Source

Highlight. A pivotal characteristic of our research is that we use data pertaining to a real system. Indeed, the webpages included in our questionnaire represent real phishing webpages that are encountered "in the wild" by real users, and which manage to bypass one of the components of a commercial-grade phishing website detector reliant on deep learning. We reached out to a large security company (which we refer to as "Sigma") whose services entail phishing website protection. In particular, Sigma employs diverse defensive mechanisms that work in tandem to minimize the chance that users fall for potential phishing "hooks". Among these, Sigma also provides a detector that leverages state-of-the-art techniques based on visual similarity (e.g., [8, 27]) to identify phishing websites.

Fig. 2: The architecture of the PDS deployed by Sigma, used as basis for the phishing examples to include in our user-study.

Phishing Detection System. The PDS used by Sigma seeks to identify phishing websites attempting to impersonate known brands (e.g., PayPal). Such a PDS (a schematic is in Fig. 2) processes diverse streams of URLs for which it tries to infer whether the corresponding website W is legitimate or not. This is done by using various DL models to compute the visual similarity between W and the entire dataset of websites associated to the brands tracked by Sigma (i.e., a given brand may have more than one website): if W is found to be "identical" to any website included in such dataset, then it is flagged as benign (and, potentially, further analysed by other mechanisms [27]). Otherwise, the decision of the PDS depends on the confidence C computed by the DL models for each website in the dataset of brands. Specifically, the PDS takes the top value of C (which is related to a specific brand B), and then compares such C against a given threshold. If C is "high", then W is considered as phishing (i.e., W is trying to impersonate B); if C is "low", then W is considered as benign (i.e., W is very different from B, and it may be a website of an unknown brand); if C is in-between, then W is not marked straight away as malicious (to avoid raising potential false positives), and is then put in a dedicated queue meant to be manually inspected by security operators. This is because Sigma seeks to improve their services by having samples that are "difficult to classify" to be used as basis to develop more robust detectors (i.e., an active-learning approach [22]).[3]

Adversarial webpages. After finding an agreement with Sigma (which required an NDA), we were given hundreds of webpages (as screenshots—in full HD resolution and taken in late 2022) which fell in-between the "high" and "low" confidence. Hence, such webpages had undergone manual verification by security analysts who verified that all such webpages were phishing websites disguised as benign webpages—thereby representing "false negatives". Interestingly, most of these phishing webpages were "poorly crafted": most humans would probably be able to suspect that "something is phishy". However, there were also cases in which the webpages exhibited a remarkable similarity with the webpage they were attempting to mimic (we provide some examples in Figs. 3 and 10). For the sake of our RQ and given the exploratory nature of our study, we considered samples from both categories. In the remainder, we will use the term "adversarial webpage" (AW) to denote a screenshot of a phishing webpage that bypassed the stricter threshold of the PDS deployed by Sigma and that was hence displayed to end users.

[3] Sigma aims to improve their PDS by having W be checked also against "known malicious examples" – triggering an immediate phishing response.

B. Questionnaire

Goals. After receiving the AW from Sigma, we devised the questionnaire used to answer our RQ. In doing so, we adhered to the following design goals (G is the motivation):

- **Heterogeneous sample** (G: anybody was eligible to participate in our user study). Since we consider AW "in the wild", anyone can land upon them. Hence, given that we are the first to conduct such a study, we follow an exploratory approach and do not set any constraint in terms of, e.g., technical expertise. We will, however, ask participants to provide some background information to enable fine-grained analyses.
- **No priming** (G: participants should not be aware that the questionnaire is related to phishing). The main reason why phishing attacks succeed is that users are distracted, or do not suspect that a given webpage may be malicious [56]. To investigate the real effectiveness of AW, we will not mention any term that may alert our participants.
- **Brand knowledge** (G: we ensure that our participants are familiar with the (legitimate) websites "mimiced" by our AW). To provide a significant answer to our RQ, we must focus on users who "can" be phished by a given AW (and inquire about their knowledge of IT).

Finally, our research is mostly tailored for Europe (due to Sigma's main location). Hence, our participants and questionnaire are going to reflect this side of the World. The list of the brands we considered (and supporting evidence) is in Table II.

TABLE II: Brands included in our questionnaire.

Brand	Category	Reason (and source)
Netflix	Video Streaming	In Q4 2022, the Europe, Middle East, and Africa demonstrated the highest concentration of paying customers for NetFlix (over 76M are from Europe) [57].
Amazon	eShop	Amazon operates in eight European countries (Germany being the most active, with €32B in net sales in 2021) [58].
Zalando	eShop	Zalando is a prominent fashion and lifestyle platform in Europe, experienced a 6% increase in active customers, surpassing 51M individuals in 2022 [59].
Airbnb	Travel	With 1.34M hosts, Europe is the largest AirBnB community worldwide [60].
Google	Information and Email	Google garners 89.3 billion monthly visits, while its email service, Gmail, enjoys widespread adoption across multiple European countries [61, 62].
Instagram	Social Network	In 2022, Europe stands as the second largest community of Instagram users, comprising an impressive population of 338 million individuals [63].
Facebook	Social Network	According to Meta, in Q4 2022 Facebook recorded 411M monthly active users in Europe (4 more than in Q3) [64].
LinkedIn	Social Network	LinkedIn, encompassing an extensive user base of nearly 1B individuals worldwide, has garnered traction within Europe, boasting a count of 242M users [65].
PayPal	Banking	PayPal revealed that its user base in Europe amounts to 35M individuals [66].
Uber	Mobility	Uber maintains a presence in a significant number of European countries [67].
Yahoo	Information and Email	Based on the Alexa rankings, Yahoo.com attains the eleventh position among the most frequently accessed websites on a global scale [68].
Twitter	Social Network	Despite being less popular than in the previous decade, Twitter is still popular in Europe (70M active users in 2023) [69].

TABLE III: Sequence of screenshots in our questionnaire, and their difficulty level. The number points to the image (hosted in our repo [13]).

#	Brand	Difficulty	Comment
1	Instagram	Hard	Resembles the legitimate login page, with the sole distinction being the footer's style.
2	Facebook	Moderate	Appears similar to the authentic version; however, suspicion may arise due to the multiple profiles that have recently logged in from the same device (specifically, six different profiles).
3	Facebook	Hard	Closely resembles the original, with the sole exception of a missing footer.
4	Instagram	Hard	Extremely challenging to distinguish, as it perfectly mirrors the original.
5	PayPal	Hard	Resembles the authentic site very closely.
6	Google	Hard	Resembles the authentic site very closely.
7	Amazon	Moderate	Resembles the authentic site very closely, but some elements have a different style.
8	Airbnb	—	It is the legitimate website.
9	Zalando	—	It is the legitimate website.
10	Netflix	Moderate	The website's header and logo may induce suspicion due to their uncharacteristic design.
11	Yahoo	Moderate	Resembles the authentic site, but some elements are stretched.
12	Yahoo	Hard	Resembles the authentic site very closely.
13	Netflix	Easy	The font style noticeably deviates from the one typically used.
14	Uber	Easy	The appearance of Uber's sign-in page notably diverges from the expected layout.
15	PayPal	Moderate	The background color of the input fields clashes with the overall design aesthetic of the website.
16	Uber	Easy	The appearance suggests it might be an outdated version of Uber.
17	LinkedIn	Easy	The font style significantly deviates from what one would expect on a professional website, disrupting its overall look and feel.
18	Netflix	Very easy	No resemblance to the original sign-up page, with a starkly contrasting and distinctive styling.
19	Twitter	Moderate	It gives the impression of being an older version of Twitter, which could still potentially elicit trust from unfamiliar users.
20	Amazon	Moderate	While it bears a striking resemblance, participants might grow suspicious due to the button on the page appearing incongruous with the overall design.

Design. To reach our goals and allow answering our RQ, we created a semi-structured questionnaire [70], which enables both qualitative and quantitative analyses. Our questionnaire is divided into three parts (in parentheses is the motivation):

I) **Demographics.** We ask preliminary questions, such as age, gender, and country of residence; but also education, expertise with IT and familiarity with some popular brands in Europe (). We need this (non-PII [71]) data to carry out fine-grained analyses, but also to comply with our third design goal, and with the European laws [72] (e.g., we manually deleted responses from people who were too young).

II) **Agreement.** We ask 20 closed questions having a similar format. Specifically, in each question we show the screenshot of a webpage, and then ask the participant to answer a question with the following phrasing: "In the screenshot, it is shown a webpage of a popular brand. Do you agree with this statement?", to which the user could respond in a 1–5 Linkert scale (with 1="Disagree", and 5="Agree"). () To comply with our second design goal, we try to avoid raising suspicion and ask a "neutral" question. Intuitively, if the participant agrees (i.e., answers with 4 or a 5), then it means that they believe the webpage to be genuine. We provide an exemplary question of part II in Fig. 3.

III) **Explanation.** We ask 20 open questions, containing the same webpages shown in the previous part. Specifically, we ask the users to explain "why" they disagreed (if so) with the statement written for the corresponding webpage (). These questions are meant to investigate what made users "suspicious" of a given webpage. This is important to determine if there are any phishing elements that are noticeable by humans, but imperceptible to ML models.

The questionnaire ended with with a last, binary question, asking whether the participant "changed their mind" about some of the answers given in the second part. We did not impose any time limit to answer any of our questions (because, in reality, users do not have such a constraint), albeit we invited not to spend more than 15s for questions in part II, and 1m for those in part III (a similar "soft-timer" was used also in [32]).

Fig. 3: Exemplary question (i.e., the first) in part II of our questionnaire. The screenshot refers to an adversarial webpage.

Content. Our semi-structured questionnaire inquires the participants' opinion on 20 different webpages (in the form of screenshot) related to 12 brands popular in Europe. Of these

20 webpages, 2 are legitimate, which we included as a form of control (which we took by ourselves in early 2023); whereas the remaining 18 are AW. Importantly, the distribution of the screenshots in our questionnaire is not fixed: such a choice is to ensure consistency, but also to further avoid priming. Indeed, we put the AW that are more likely to raise suspicion (due to, e.g., clearly different logos) at the end of the questionnaire: intuitively, if participants were shown suspicious webpages from the beginning, then they would be more skeptical of the remaining webpages—thereby leading to biased results. The 18 AW, as well as their placement in our questionnaire, were chosen after many meetings (some of which included Google employees), in which the authors discussed the peculiar characteristics of each AW and eventually reached a consensus on a qualitative "difficulty level" to identify an AW as phishing. We report in Table III the sequence of our screenshots, their brand and their difficulty level.

Data collection. We shared our questionnaire on popular social media (e.g., LinkedIn).[4] To avoid priming, we advertised it as "Website Content Perception Survey". We did not provide any payment to participants. We began collecting answers[5] on May 18th, 2023; and stopped after three weeks. We provide our complete questionnaire in our repository [13]. Ethical remarks are discussed in §VIII.

C. Sample and Limitations

We describe our sample and the limitations of our research.

Sample Description. We received 126 responses. Gender-wise, 70 (55.6%) identified themselves as male, and 54 (43.6%) as female (1 did not answer). In terms of age, 3 (2.4%) are younger than 16; 44 (34.9%) are in the 16–24 range; 57 (45.2%) between 25–34; 12 (9.5%) between 35–44; 4 (3.1%) between 45–54; 6 (4.8%) between 55–64; and none are older than 65. With regards to country of residence, 70 (61.9%) are from Austria; 19 (15.1%) are from Germany, and 12 (9.5%) from Switzerland; 5 (3.9%) are from Bosnia, 4 (3.2%) are from Slovenia, and 2 (1.6%) from Liechtenstein; 1 (0.8%) are from Finland, Georgia, Macedonia, Estonia, Italy, as well as from the USA. The educational background of the participants accentuated the diverse nature of our sample: 45 (35.7%) have a high-school diploma, whereas 42 (33.3%) a BSc, and 27 (21.4%) an MSc; 2 (1.6%) have a PhD; 11 (8.7%) only completed basic schooling. With respect to expertise with IT, 75 (59.5%) participants are heavily involved with IT (either professionally or for personal interests), whereas 48 (38.1%) only use IT for entertainment or when necessary; lastly, 3 (2.4%) reported to make a very limited use of IT in their daily lives. Finally, we report in Table IV the amount of participants that are familiar with the brands we considered, showing that most of our sample is familiar with our chosen brands—validating the real-world applicability of our findings.

TABLE IV: Familiarity of our sample with the twelve brands in our questionnaire. On average, our brands are known by 91 (72%) participants.

Brand	NetFlix	Amazon	Zalando	AirBnB	Google	Instagram	Facebook	LinkedIn	PayPal	Uber	Yahoo	Twitter
Absolute	117	109	92	80	123	117	98	96	93	55	42	68
Relative	93%	87%	73%	63%	98%	93%	78%	76%	74%	44%	33%	54%

Limitations. Our sample and questionnaire have some intrinsic limitations. For instance, most of our participants are from German-speaking countries, so our study is biased towards this area. Furthermore, our sample exhibits significant diversity in terms of age, gender and background: this characteristic is both a strength (since it allows deriving broad takeaways) and a weakness (since it impairs specificity). Finally, our questionnaire includes only 18 AW pertaining to some reputable brands: therefore, there may be other brands, or other types of AW, for which our study cannot provide any answer. For these reasons, we do not claim that the results of our research can be generalized to reflect the entire human population[6] and/or the full landscape of phishing threats. Nonetheless, given the exploratory nature of our study (which is the first to investigate our RQ) as well as the many precautions we took to reduce priming and ensure realistic assessments, our results represent a significant step towards mitigating the proliferation of phishing websites.

IV. Results (Quantitative)

We now report the quantitative results of our user-study (from part II). We first show high-level findings (§IV-A), and then focus on specific subsets of our sample (§IV-B). We also analyze two intriguing phenomena on the natural progression of our questionnaire (§IV-C). Finally, we perform fine-grained analyses on some relevant combinations of our sample's demographics (§IV-D). While presenting the results, we will make some "claims" (denoted as [C]), which we validate through statistical tests in the next section (§V-A).

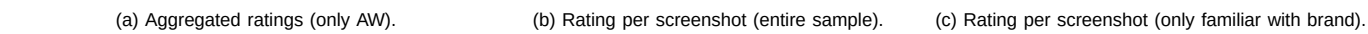
A. General Findings

We begin by addressing our main RQ at a high-level. We report in Fig. 4a the distribution of the "agreement ratings" that our participants provided for all the 18 AW screenshots in our questionnaire. (As a reminder, higher agreement implies higher likelihood of being deceived.) Specifically, we provide three boxplots: the leftmost one represents our entire sample (having 2268 ratings, given by 126 participants * 18 AW); the central one represents the ratings provided only by those participants who reported being "familiar" with the corresponding brand of each AW (having 1666 ratings); whereas the rightmost one represents the ratings of those participants who stated otherwise (having 602 ratings). From Fig. 4a, we can see that the ratings of most of our sample are above the saddle point (of 3), and the mean for these three boxplots is 4. These results suggest that our chosen AW deceive the

[4] We provided the questionnaire both in English and in German to enable even people with limited knowledge of English to participate.

[5] We conducted pilot tests with colleagues. After submitting their responses, we revealed that 18 out of 20 screenshots were phishing. These pilot participants did not expect this, and some stated to be "embarassed" for being unable to figure this out.

[6] We also observe that our sample is larger than the one considered by most prior work (c.f. Table I), and that even recent top-conferences accepted papers (e.g., [73]) having user-studies with a smaller population than ours.



(a) Aggregated ratings (only AW). (b) Rating per screenshot (entire sample). (c) Rating per screenshot (only familiar with brand).

Fig. 4: High-level results. Fig. 4a reports the aggregated distribution for the 18 AW in our questionnaire. Figs. 4b and 4c show the rating distribution for each screenshot (green for legitimate, red for AW). Our AW can deceive most users (especially at the start of the questionnaire).

(a) Education. (b) Gender. (c) Expertise with IT. (d) Age.

Fig. 5: Subgroup results. The figures report the aggregated ratings (for the 18 AW) of each subgroup (the x-axis denotes the size of each subgroup).

human users. Surprisingly, it appears that users who are more familiar with a given brand tend to be easier to deceive [C1].

To provide a deeper understanding, we report the ratings provided to each individual screenshot (AW are in red, whereas legitimate websites are in green) by our entire sample (Fig. 4b) and by only those who are "familiar" with the brand of the screenshot (Fig. 4c). We see that "familiar" participants are worse at identifying the phishing nature of screenshots #4, #13 and #17. Finally, we see that the rating for the first five AW is very high (around 4.5 on average) meaning that these AW can successfully "phish" most of our sample—under the assumption that (i) no additional contextual information is provided and that (ii) there is no priming.

> TAKEAWAY. Most of our sample cannot recognize AW, and familiarity with a brand hinders the detection skills of users.

B. Group-specific analyses

We now focus our attention on specific subsets of our sample. We rely on the demographics information provided by participants at the beginning. We do so by providing the aggregated rating distribution (only for the 18 AW) of our sample on the basis of: education (Fig. 5a), gender (Fig. 5b), expertise with IT (Fig. 5c), and age (Fig. 5d); the x-axes of all figures in Fig. 5 report the population of each subgroup.

By observing Figs. 5, we can make four significant claims:

[C2] University graduates are more suspicious (Fig. 5a).

[C3] Female appear to be less suspicious than males (Fig. 5b).

[C4] IT experts are more skeptical than amateurs (Fig. 5c).

[C5] Age is not correlated with suspiciousness (Fig. 5d).

Regardless, we found it intriguing that those who possess a PhD tend to be the easiest to be deceived—albeit we cannot support such a claim, since only two participants of our sample have a doctorate.

C. Skepticism over-time

We find it instructive to analyze the natural progression of the ratings as each participant advanced in the questionnaire. Indeed, we recall (§III-B) that – to avoid priming – we placed the hardest AW to identify at the beginning of the questionnaire, whereas the easiest were at the end. Hence, our participants are bound to become more suspicious over-time, which would lead to a drop in the agreement ratings. We perform this exercise by focusing on two demographics: gender (Figs. 6) and expertise with IT (Figs. 7).

Male vs Females. By comparing Fig. 6a with Fig. 6b, we can see that both groups exhibit high agreement (avg 4.3 for males, and 4.5 for females) in the first 7 screenshots—all being AW. (Interestingly, males tend to be dubious of the two legitimate screenshots!) However, starting from 10th screenshot, the agreement of males starts decreasing (avg 3.7), whereas those of females remains high (avg 4.2) until screenshot #18, which leads to a drop in agreement by most (avg 2.8 for males, and 3.3 for females).

(a) Male (N=70). (b) Female (N=55).

Fig. 6: Individual screenshot ratings based on Gender.

Experts vs Amateurs.[7] By comparing Fig. 7a with Fig. 7b, we see some interesting trends. Specifically, "experts" tend to agree similarly to "amateurs" at the beginning [C6]; however, after completing half of the questionnaire, "experts" become much more skeptical than "amateurs" [C7].

(a) IT experts (N=75). (b) IT amateurs (N=48).

Fig. 7: Individual screenshot ratings based on Expertise with IT.

TAKEAWAY. As participants advance in our questionnaire, they appear to become more suspicious.

D. Fine-grained analyses

We conclude our results by focusing our attention at various subgroups of our sample, which we draw from those having the highest amount of participants. Specifically, we consider the individual ratings of participants who are aged 16–24 (Figs. 8) or 25–34 (Figs. 9) and that are either male or female, and either IT experts or amateurs—thereby resulting in eight different combinations. While we acknowledge that some of them have few examples, we find it educational to analyze these case-studies—which can be considered extensions of those discussed in §IV-C,

By observing Figs. 8 and Figs. 9, we derive the following:

- Female IT amateurs become less skeptical as they age (cf. Fig. 8d with 9d)...
- ...but the opposite holds for males (cf. Fig. 8c with 9c).
- Legitimate webpages appear suspicious to male IT amateurs aged 25–34, but not for 16–24 (cf. Fig. 9c with 8c)...
- ...but the opposite holds for male IT experts: those aged 16–24 are more suspicious of legitimate webpages w.r.t. those aged 25–34 (cf. Fig. 8a with Fig. 9a).
- Female IT amateurs aged 25–34 have the highest "agreement" ratings—making them more susceptible to AW.

Given the small sample size, we refrain from making claims on the these observations. However, the lesson learnt is that some groups of users are more vulnerable to AW than others.

V. VALIDATION AND ANALYSIS

We expand our quantitative analyses on part II. We validate our claims (§V-A) and draw similarities with prior work (§V-B).

A. Statistical validation

We validate our 7 claims made in §IV through statistical tests.

Hypotheses. Inspired by prior work [6], we rely on the Welch's t-test [74]. This test can determine if two groups are equal by comparing the resulting p-value with a given target (typically set to 0.05). Hence, for each claim, we identify two groups ($g1$ and $g2$), compute the p-value, and use it to test the null hypothesis (H_0): "$g1$ and $g2$ are statistically equivalent". H_0 is accepted if $p > 0.05$, and rejected otherwise. It can also be that the test is inconclusive (due to, e.g., lack of data-points): to provide more confident conclusions, we also measure the effect size (ES) of each test.

Setup. All our claims refer to how our participants analyzed phishing screenshots. For each test, our groups entail the agreement ratings provided by the two compared groups ($g1$ and $g2$) to a specified set of AW screenshots. Let us identify the groups we considered in our tests to validate each claim.

- C1: Familiarity. (H_0 should be rejected.) $g1$ denotes participants who are familiar with the brand of a given AW, whereas $g2$ denotes those who are not familiar.
- C2: University. (H_0 should be rejected.) $g1$ denotes participants with a degree (BSc., MSc., PhD), whereas $g2$ are those without a degree (Basic school or high-school).
- C3: Gender. (H_0 should be rejected.) $g1$ denotes those who identified as male, and $g2$ as female.
- C4: IT expertise. (H_0 should be rejected.) $g1$ denotes those who are "experts", and $g2$ "amateurs".
- C5: Age. (H_0 should be accepted.) $g1$ denotes participants aged < 25 (47 in total), $g2$ those aged 25–34 (57); we also consider $g3$ including those > 34 (22).
- C6: Similar beginning. (H_0 should be accepted.) We consider the first 4 AW; $g1$ denotes experts in IT, and $g2$ amateurs.
- C7: Different ending. (H_0 should be rejected.) We consider the last 10 AW; $g1$ denotes experts in IT, and $g2$ amateurs.

For [C1]–[C5], $g1$ and $g2$ include all 18 AW.

Results. We display the results of these tests in Table V, in which rows report the amount of elements (N), the average and standard deviation of each group; as well as the p-value (green/red cells denote cases in which H_0 must be accepted/rejected) and the ES of the test. Table V shows that our claims are validated: cases in which H_0 must be rejected also show a small ES, which provides additional evidence that the two groups are statistically different. Finally, for C5 (for which

[7] We use "amateur" to denote participants who "use IT only when necessary or for entertainment", and "expert" for those who are "passionate about IT".

(a) Male, IT Experts (N=11). (b) Female, IT Experts (N=12). (c) Male, IT amateurs (N=7). (d) Female, IT amateurs (N=12).

Fig. 8: Case-study. Individual screenshot ratings of participants aged 16–24 (N=44), categorized on the basis of gender and IT expertise.

(a) Male, IT Experts (N=28). (b) Female, IT Experts (N=7). (c) Male, IT amateurs (N=11). (d) Female, IT amateurs (N=10).

Fig. 9: Case-study. Individual screenshot ratings of participants aged 25–34 (N=57), categorized on the basis of gender and IT expertise.

we identi ed 3 groups), we also compared $g2$ with $g3$ (having avg=4:07, std=1:13, N=396) and we found that $p=0:42$ (ES=0:047) i.e., H_0 must be accepted (since $p > =0:05$); hence, since $g1 \ g2$, and $g2 \ g3$, it follows that $g1 \ g3$, which validates the claim that age has a negligible impact on phishing awareness—at least according to our sample.

TABLE V: Statistical validation of our claimed hypotheses ($=0:05$)

Claim	C1		C2		C3		C4		C5		C6		C7	
	g1	g2	g1	g2	g1	g2	g1	g2	g1	g2	g1	g2	g1	g2
N	1666	602	1260	1008	1260	990	1350	864	846	1080	300	192	750	480
avg.	4:14	3:96	3:94	4:29	3:92	4:30	3:98	4:27	4:13	4:07	4:46	4:56	3:65	4:10
std.	1:31	1:29	1:38	1:18	1:43	1:10	1:41	1:13	1:22	1:41	1:03	0:88	1:54	1:23
p	0:004		< 0:001		< 0:001		< 0:001		0:32		0:26		< 0:001	
ES	0:14		0:27		0:29		0:22		0:046		0:1		0:31	

Remark: All our claims and findings pertain to the data of our user-study. We do not generalize (see §III-C).

B. Comparison with prior work

We appreciate that our sample bears some resemblance with the one of prior studies [6]. For instance, our gender distribution is similar to the user-studies in [7, 37, 49]. Interestingly, Orunsolu et al. [7] found that female perform better (their sample had a male:female split of 57:43), while the study by Kumaraguru et al. [44] found otherwise (albeit the 28 participants in [44] had a male-to-female ratio of 5-to-1). Despite having a different goal in mind, our findings align with those in [44] (see C3).

From an age perspective, Purkait et al. [47] (whose sample was spread between 20–62 years of age) found that elders were more susceptible; the opposite was found by Kumaraguru [44] (whose sample was aged 13–65), who claimed that participants between 13–17 were the most susceptible to phishing—a finding shared also by Lastdrager et al. [37]. In contrast, we did not find any significant performance difference related to age (see C5)—a result that aligns with those in [40, 42, 54]. Finally, the user-studies in [40, 49] found that expertise with IT was not correlated with phishing susceptibility. In contrast, Orunsolu et al. [7] found that experts are more resilient—which aligns with our C4 (albeit this may not hold for specific subgroups §IV-D). We stress, however, that drawing conclusions based on similar correlations may be misleading—as echoed in a very recent work [75].

Remark: our user-study has different goals than those by prior work (§II-C). Hence, comparisons may not be appropriate, and we do these solely for educational purposes.

VI. Explanations (qualitative)

We qualitatively analyze the responses we received for part III.

A. Considered screenshots

Our participants provided plenty of (unstructured) comments in part III, and objectively analyzing all of these is impossible—given that such responses are also in diverse languages (which would further add bias in the translation). Hence, we prefer to focus on the responses we received for three meaningful screenshots, namely:

Screenshot #1 (Instagram—hard difficulty, shown in Fig. 3)) This is the first screenshot of our questionnaire. Hence, it represents the perfect use-case since there is no form of “phishing priming” that may influence the

agreement of our participants. Perhaps unsurprisingly, the average rating for this screenshot is 4.8 (cf. Fig. 4b).

Screenshot #10 (Net ix—moderate dif culty, cf. Fig. 10a).) This screenshot is at the middle of our questionnaire, and represents a good balance between dif culty (it is easier to identify than #1) and priming (participants may have begun to become suspicious after answering the previous nine questions). Its average rating is 4.5 (cf. Fig. 4b).

Screenshot #18 (Net ix—very easy dif culty, cf. Fig. 10b).) This screenshot is at the end of our questionnaire, and aside from being very easy to identify as phishing, it also refers to Net ix (same as #10). Analyzing #18 is useful to investigate what elements of an AW are "easily spotted" by humans, thereby allowing practitioners to either ignore similar AW (since humans can easily recognize them as phishing) or focus on them (to avoid annoying users) to improve their PDS. Its average rating is 3 (cf. Fig. 4b).

To avoid bias due to translation, we base these analyses on the responses of German-speaking participants, which represent 82% (i.e., 103 out of 126) of our sample (see §III-C).

B. Assessment

Overall, 8 (8%) rated screenshot #1 with a 3 or less, and 18 (17%) wrote a comment on the corresponding question in part III; for screenshot #10, 16 (16%) rated it with a 3 or less, and 21 (20%) wrote a comment on it; for screenshot #18, 61 (59%) rated it with a 3 or less, and 69 (67%) wrote a comment.

Let us report some remarks written by our participants.

#1: the few comments "con rm" that participants agree with the statement. Others reported "not having enough knowledge about Instagram to con rm certain statements". The only valuable remarks from a phishing perspective are those by users who reported that the presentation of the webpage is "incorrect", and that there is a "lack of the logo" (which we found to be odd, since screenshot #1 has the Instagram logo, and even the real login webpage of Instagram does not have another logo in it).

#10: many participants expressed concerns on the logo, which is described as being "distorted" or "inauthentic". Some mentioned weird placement of "tabs", and that the screenshot lacks a "search function". Few mentioned that NetFlix does not offer only "movies", but also "documentaries" (not shown in the screenshot). The "font type and "headings" were also mentioned as source of dou

#18: many commented that the logo is "incorrect" or "outdated". Concerns were made on the overall look of the webpage, which appears "cheap", "unprofessional" and "untrustworthy". Speci cally, some stated that it "does not say anything about Net ix" (i.e., there are no "images or movies") and that it only resembles a "registration page". The lack of a "search function" was also reported frequently. Some participants also criticized the "colo scheme", which does not match the one of Net ix.

Given the abundant feedback we received for screenshot #10 and #18, we visualize these comments (in Fig. 11) by making a word cloud (in German—the English translation is in Table VII) of the responses we received. To protect the privacy of our participants, we cannot report the verbatim German text.

> Remark: the lack of comments for screenshot #1 (and its high agreement rating of 4.8) is evidence that it deceived most users, but also shows that our questionnaire resembled a realistic phishing scenario wherein users are not primed.

C. Interpretation (with practitioners)

Insofar, we have provided generic remarks that our participants expressed on these three screenshots. We now attempt to elaborate actionable insights—with the assistance of Sigma Coding. We held six meetings with Sigma's employees, focused on performing inductive coding sessions [76]. The goal was to devise a codebook used to identify which (visual) elements in a screenshot of an AW can be used to infer that the corresponding image relates to a phishing webpage. Intuitively, by (i) identifying such elements, and then (ii) quantitatively measuring their prevalence "in the wild", it would be possible to determine which aspects should be prioritized by practitioners to improve the their PDS. During a meeting, the attendees discussed many screenshots of AW, attempting to derive an "actionable" set of phishing elements. After six meetings, the codebook encompasses 9 elements. Among these, we cite: "altered visual logo", "different style of text and font", and "unusual login functionality and style".

Mapping. We nd it instructive to use the feedback received by our participants, and "map" it to these three abovementioned elements. This is useful as a form of validation: "do netizens also see the same elements that we see?". Indeed, if these elements appear to have been noticed also by other internet users, then they can be acted upon to improve the detection mechanisms of PDS so that they can better deal with evasive phishing webpages. Due to the few comments received for screenshot #1, we will only do this mapping for screenshot #10 and #18. We report in Table VI our translation (which does not breach privacy) of those statements (in random order) that can be mapped to these three elements we identi ed. We recall that questions in part III asked participants to explain "why did you disagree with the statement in part II?". We could not nd any statement for "unusual login functionality" (since #10 does not have it in the rst place—see Fig. 10a).

> TAKEAWAY . Several participants noticed some "common phishing elements" that can be acted upon (by practitioners) to improve existing PDS against (real) evasive webpages.

Countermeasure. Based on these explanations, Sigma is currently working towards a solution that is better equipped to counter similar phishing webpages—which can be somewhat detected by real users, but which are still an annoyance. Furthermore, an orthogonal objective pursued by Sigma is to identify some elements of AW that deceived most human users (e.g., #1 ; #2 ; #3) and develop appropriate countermeasures.

TABLE VI: Mapping of participants' explanations (for Screenshots #10 and #18) to three of our codes. (Screenshot #10 does not have any login form)

	Altered Visual Logo	Unusual Login Functionality and Style	Different style of text and font
Screenshot #10	"because of the logo. It's squeezed together" "logo/branding looks fake. The font on the categories doesn't t." "Logo is not on top right and everything is very distorted/compressed" "Looks fake. (Logo, layout)" "slightly different logo"	N/A	"Looks a little distorted in the picture, not sure. May well be fake" "weird rendering and font" "Logo, Layout" "The interface of Net ix looks different. The "tabs" are arranged on the left, etc." "Wasn't exactly sure-the headings look different somehow (font & size)."
Screenshot #18	"wrong Net ix logo - fake" "wrong logo, it hasn't existed like this for years" "wrong logo" "I nd the logo weird, but it seems to be the page for registration, so not login but registration if the logo is not fake" "different logo and different colors" "completely different logo"	"Screenshot looks more like password renewa" "completely different interface, Net ix doesn't use blue as much, generally different login and design" "the Net ix login page looks different in my opinion" "you can see the registration page not the login page" "the login page looks different than what I'm used to. I nd a little confusing/different" "not login, but password change" "the registration page of Net ix that I know looks different"	"modern login page looks different" "too minimalistic if you don't know the site" "looks cheap, something is wrong there" "Layout is too old fashioned, today Net ix login looks different" "looks like a fake site" "outdated design"

Nonetheless, we are currently working with Sigma to quantify the prevalence of "elusive elements" in AW, which can be used as a guide for practitioners to determine which elements are more common and hence should be given priority.

"Would you change your mind?": Recall that our questionnaire ends with a binary question asking whether a given participant was willing to change their initial ratings if given another opportunity (§III-B). Out of our 126 participants, 92 (73%) af rmed that they would not change their mind, whereas 34 (27%) stated otherwise.

VII. Discussion

We discuss some alternative ways to carry out our study (§VII-A). Then, as a nal contribution of our paper, we draw actionable recommendations for related research (§VII-B).

A. Alternative formulations

We discuss four speci c design choices of our questionnaire.

Structure of the questionnaire. A valid point (which was also raised during our pilot study) is that asking the users to explain their disagreement after having completed part II can lead to users "forgetting" why they disagreed with some statements. We acknowledge such a remark: however, we did this because our main focus is determining if users are tricked by AW – which is addressed by part II. Asking the users to provide an explanation immediately after answering could have increased their suspiciousness, leading to less realistic responses for part II (which is our main focus) in favor of more details for part III (which relates to a relevant, but ancillary problem).

Phrasing of the questions in part II. Among our priorities was to minimize the amount of priming[8], which is why we opted for a neutral (and, potentially, vague[9]) question to be asked in part II. Of course, we could have asked "do you think that this screenshot represents a legitimate webpage?" (similarly to, e.g., [49]): however, doing so would have led our participants to be suspicious of every webpage—which is not realistic in a phishing context, given that phishing is successful when users do not expect it (which also explains why most employees get phished despite receiving proper education [78, 79])[10].

Absence of context. In our study, users are not given any information about "why" they would land on a given webpage. For instance, in a real setting, a user may be shown a webpage after clicking on a link (received, e.g., via email or instant messaging)[11]. We acknowledge that context can be an important source to determine whether a website is phishing or not; however, our design choice is appropriate to answer our RQ, whose goal is to investigate the susceptibility of users to AW, i.e., phishing webpages that evaded a PDS. If a user becomes suspicious of a webpage "because of context" then it would be unfair to the PDS (which, ttbook, do not account for context—yet). Furthermore, users who are suf ciently alert to become suspicious due to context are also less likely to fall for phishing in the rst place [6]: hence, lack of context can be seen as a scenario in which users do not suspect anything—which are the most dangerous, from a phishing perspective.

Number of AW. Our questionnaire entails 18 AW (which are xed[12] for every participant), but Sigma provided us with a much higher number. While we acknowledge that we could have included more AW in our study, we did not do so for two reasons. First, because adding more AW to our questionnaire would have increased its length, thereby: decreasing the level of attention of each participant; increasing the suspiciousness of each participant for any additional question; and potentially discouraging more users to participate (while part II was took 5 minutes to complete, part III took 15 minutes). Second, because having each participant provide their opinion on a different set of AW would have prevented one from analyzing trends about individual AW (such as, e.g., investigating which AW tricked most users, and trying to understand "why").

Simply put, there are many ways in which our study could have been designed—each with its pros and cons. Our choices are driven by our primary goals, dictated by our main RQ.

B. Recommendations for Research

Let us coalesce all our ndings and derive recommendations for researchers. First, we endorse "technical papers" on phishing website detection to embrace our overarching message: carry out user-studies that focus on investigating how real

[8] Designing bias-free user-studies for phishing is an open problem [49, 77].

[9] We never use terms such as "trust", "malicious", "legitimate", "phishing"

[10] An interesting question to ask at the very end of our questionnaire is "Did you gure out that this questionnaire was about phishing awareness (and if so, when)?", which would have acted as additional validation.

[11] In our questionnaire, we provided screenshots as rendered by a desktop Web-browser, hence we cannot assess the impact of phishing on mobile devices (e.g., smartphones, tablets). We encourage future research to do so.

[12] A random ordering could have been useful to, e.g., ascertain whether the skepticism over time is truly caused by the natural progression of the exercise, or by the different dif culties of the screenshots. However, random ordering would have also prevented a fair comparison for other effects that were more important for the sake of our study (see §III-B).

users perceive the corresponding phishing websites.[13] Then, we make three observations, rooted on our own experience and findings, that can help devising meaningful user-studies.

It's feasible. As our study showed, carrying out such user-studies is tough, but not impossible. Ultimately, we devised a questionnaire, advertised it on popular social media, and analysed the responses we collected over three weeks. Alternatively, the recent work by Lee et al. [32] relied on Amazon Mechanical Turk. Such an additional validation would dramatically increase the real-world value of the findings of a research paper.

Avoid priming. Users are more skeptical of webpages when they are aware that they may be concealing a phishing trap—which may bias results. Hence, we recommend that future user-studies refrain from priming users.

Make it short. An important finding of our study is that, even when users are not primed, they may naturally become more suspicious of the samples shown during a questionnaire—if such samples exhibit strong elements that "something is phishy." Hence, we recommend that (i) future user-studies only show few samples to any given participant, and that (ii) account for the fact that the responses for the last samples may be biased (due to the natural priming).

Finally, we remark that future efforts can even use our template (which we release [13]) as basis for their questionnaires.

VIII. Conclusions and Future Work

Countering phishing websites is a two-step decision process, entailing both "machines" (which provide a first layer of defense) and "humans" (who are the true target). Unfortunately, in research, prior work mostly focused on either one of these steps. At the same time, in reality, existing phishing detection systems (PDS) cannot detect all phishing websites, and end-users still fall for phish.

In this paper, we advocate to change the panorama of anti-phishing schemes in research. We do so by linking the response of humans with that of (real) machine learning-based PDS. We hope future endeavours will embrace the direction of our work. For instance, researchers can assess the response of humans to their proposed PDS, thereby pinpointing which phishing techniques can simultaneously deceive both machines and users. The corresponding findings can then be used by practitioners to refine their operational PDS. Ultimately, perfect detection is an enticing but unattainable goal: resources should be spent on countering those phishing webpages that are more likely to trick humans.

Ethical Statement. Our institutions are aware of and approve the research discussed in this paper. The respondents to our questionnaire know the identity of the author who collected their data, and we are willing to delete their data should they ask us to do so. The participants were made aware that their data was going to be privately stored, which is why we cannot disclose the full responses. To comply with the Menlo report [80], we never asked for sensitive data (i.e., [81, 82]) or for personal identifiable information [71] in our questionnaire; moreover, we only show screenshots of phishing webpages (which are reachable by unpublished urls), and do not deploy any phishing webpage "on the web". We also complied with European regulation, and ensure that our participants are old enough to respond to online surveys without the explicit consent of a tutor (which varies from country to country [72]). Due to NDA with Sigma, we cannot disclose more information about the considered detector, the considered screenshots (which we do release in our repository [13]), the effects that these screenshots had on Sigma's customers, or on Sigma itself.

Acknowledgements. The authors would like to thank the anonymous reviewers (as well as the eCrime'23 attendees) for their feedback. We also thank the Hilti Corporation for funding, and Sigma for allowing us to carry out this study.

[13] For papers that propose "novel attacks" that bypass existing anti-phishing schemes, such user-studies should verify whether users are really deceived by the evasive webpages; whereas papers that propose "novel defenses", the focus should be on the webpages that still mange to evade the robust PDS.

References

[1] "State of the phish 2022," https://www.proofpoint.com/it/resources/threat-reports/state-of-phish, ProofPoint, Tech. Rep., 2022.
[2] M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection: a literature survey," IEEE Communications Surveys & Tutorials, 2013.
[3] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," Pattern Recognition, 2018.
[4] "Interet crime report," https://www.ic3.gov/Media/PDF/AnnualReport/2022_IC3Report.pdf, Federal Bur. of Investigation, Tech. Rep., 2022.
[5] "Phishing activity trends report," APWG, Tech. Rep., 2022, https://docs.apwg.org/reports/apwg_trends_report_q2_2022.pdf.
[6] S. Baki and R. M. Verma, "Sixteen years of phishing user studies: What have we learned?" IEEE TDSC, 2022.
[7] A. A. Orunsolu, O. Afolabi, A. S. Sodiya, A. T. Akinwale et al., "A users' awareness study and influence of socio-demography perception of anti-phishing security tips," Acta Informatica Pragensia, 2018.
[8] S. Abdelnabi, K. Krombholz, and M. Fritz, "Visualphishnet: Zero-day phishing website detection by visual similarity," ACM CCS, 2020.
[9] S. Bell and P. Komisarczuk, "An analysis of phishing blacklists: Google safe browsing, openphish, and phishtank," ACSW, 2020.
[10] G. Apruzzese, H. S. Anderson, S. Dambra, D. Freeman, F. Pierazzi, and K. Roundy, ""Real Attackers Don't Compute Gradients": Bridging the Gap Between Adversarial ML Research and Practice," SaTML, 2023.
[11] A. Oest, Y. Safaei, P. Zhang, B. Wardman, K. Tyers, Y. Shoshitaishvili, and A. Doupé, "PhishTime: Continuous longitudinal measurement of the effectiveness of anti-phishing blacklists," USENIX Security, 2020.
[12] D. M. Divakaran and A. Oest, "Phishing detection leveraging machine learning and deep learning: A review," IEEE Security & Privacy, 2022.
[13] "Our repo," https://github.com/hihey54/eCrime23_RealAdversarialPhish.
[14] B. Ampel, Y. Gao, J. Hu, S. Samtani, and H. Chen, "Benchmarking the robustness of phishing email detection systems," ACM CIS, 2023.
[15] L. Kersten, P. Burda, L. Allodi, and N. Zannone, "Investigating the effect of phishing believability on phishing reporting," in IEEE European Symposium on Security and Privacy Workshops, 2022.
[16] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs, "Who falls for phish? a demographic analysis of phishing susceptibility and effectiveness of interventions," Proc. Conf. HFCS, 2010.
[17] I. Corona, B. Biggio, M. Contini, L. Piras, R. Corda, M. Mereu, G. Mureddu, D. Ariu, and F. Roli, "Deltaphish: Detecting phishing webpages in compromised websites," ESORICS, 2017.
[18] B. Kondracki, B. A. Azad, O. Starov, and N. Nikiforakis, "Catching transparent phish: Analyzing and detecting mitm phishing toolkits," in ACM CCS, 2021.
[19] Á. Feal, P. Vallina, J. Gamba, S. Pastrana, A. Nappa, O. Hohlfeld, N. Vallina-Rodriguez, and J. Tapiador, "Blocklist babel: On the transparency and dynamics of open source blocklisting," IEEE Transactions on Network and Service Management, 2021.

[20] K. Tian, S. T. Jan, H. Hu, D. Yao, and G. Wang, "Needle in a haystack: Tracking down elite phishing domains in the wild," in IMC, 2018.

[21] Y. Zhang, J. I. Hong, and L. F. Cranor, "Cantina: a content-based approach to detecting phishing web sites," in Proc. WWW, 2007.

[22] G. Apruzzese et al., "The role of machine learning in cybersecurity," ACM Digital Threats: Research and Practice, 2022.

[23] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Predicting phishing websites based on self-structuring neural network," Neur. Comp. Appl., 2014.

[24] Q. Cui, G.-V. Jourdan, G. v. Bochmann, and I.-V. Onut, "SemanticPhish: a semantic-based scanning system for early detection of phishing attacks," in APWG Symposium on Electronic Crime Research (eCrime), 2020.

[25] A. Y. Fu, L. Wenyin, and X. Deng, "Detecting phishing web pages with visual similarity assessment based on earth mover's distance (emd)," IEEE TDSC, 2006.

[26] B. Van Dooremaal, P. Burda, L. Allodi, and N. Zannone, "Combining text and visual features to improve the identi cation of cloned webpages for early phishing detection," in Proc. ARES, 2021.

[27] Y. Lin, R. Liu, D. M. Divakaran et al., "Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages," in Proc. USENIX Secur. Symp., 2021.

[28] R. Liu, Y. Lin, X. Yang, S. H. Ng, D. M. Divakaran, and J. S. Dong, "Inferring phishing intention via webpage appearance and dynamics: A deep vision based approach," in USENIX Security, 2022.

[29] B. Liang et al., "Cracking classi ers for evasion: a case study on the google's phishing pages lter," in WWW, 2016.

[30] Y. Gao, B. M. Ampel, and S. Samtani, "Evading anti-phishing models: A eld note documenting an experience in the machine learning security evasion competition 2022," ACM DTRAP, 2023.

[31] G. Apruzzese, M. Conti, and Y. Yuan, "Spacephish: The evasion-space of adversarial attacks against phishing website detectors using machine learning," in Proc. ACSAC, 2022.

[32] J. Lee, Z. Xin, M. P. S. Ng, K. Sabharwal, G. Apruzzese, and D. M. Divakaran, "Attacking logo-based phishing website detectors with adversarial perturbations," in European Symposium on Research in Computer Security (ESORICS), 2023.

[33] A. Hutchings and T. J. Holt, "The online stolen data market: disruption and intervention approaches," Global Crime, 2017.

[34] P. Zhang, Z. Sun, S. Kyung, H. W. Behrens, Z. L. Basque, H. Cho, A. Oest, R. Wang, T. Bao, Y. Shoshitaishvili et al., "I'm SPARTACUS, No, I'm SPARTACUS: Proactively Protecting Users from Phishing by Intentionally Triggering Cloaking Behavior," in Proc. ACSAC, 2022.

[35] G. Apruzzese and V. Subrahmanian, "Mitigating adversarial gray-box attacks against phishing detectors," IEEE Transactions on Dependable and Secure Computing, 2022.

[36] C. Iuga, J. R. Nurse, and A. Erola, "Baiting the hook: factors impacting susceptibility to phishing attacks," HCIS, 2016.

[37] E. Lastdrager, I. C. Gallardo, P. Hartel, and M. Junger, "How effective is anti-phishing training for children," in SOUPS, 2017.

[38] R. Dhamija, J. D. Tygar, and M. Hearst, "Why phishing works," in Proc. SIGCHI CHI, 2006.

[39] A. Tsow and M. Jakobsson, "Deceit and deception: A large user study of phishing," Indiana University, vol. 9, 2007.

[40] S. Sheng et al., "Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish," in SOUPS, 2007.

[41] M. Jakobsson, "The human factor in phishing," Privacy & Security of Consumer Information, vol. 7, no. 1, pp. 1–19, 2007.

[42] A. Herzberg and A. Jbara, "Security and identi cation indicators for browsers against spoo ng and phishing attacks," ACM TIT, 2008.

[43] A. Alnajim and M. Munro, "An anti-phishing approach that uses training intervention for phishing websites detection," in IEEE ITNG, 2009.

[44] P. Kumaraguru, S. Sheng, A. Acquisti, L. F. Cranor, and J. Hong, "Teaching johnny not to fall for phish," TOIT, 2010.

[45] C.-C. Yang, S.-S. Tseng, T.-J. Lee, J.-F. Weng, and K. Chen, "Building an anti-phishing game to enhance network security literacy learning," in IEEE Int. Conf. Adv. Learn. Tech., 2012.

[46] N. Asanka, G. Arachchilage, S. Love, and C. Maple, "Can a mobile game teach computer users to thwart phishing attacks?" International Journal for Infonomics, 2013.

[47] S. Purkait, S. Kumar De, and D. Suar, "An empirical investigation of the factors that in uence internet user's ability to correctly identify a phishing website," Inf. Manag. & Comp. Secur., 2014.

[48] M. J. Scott, G. Ghinea, and N. A. G. Arachchilage, "Assessing the role of conceptual knowledge in an anti-phishing educational game," in IEEE Int. Conf. Advanced Learn. Tech., 2014.

[49] M. Alsharnouby, F. Alaca, and S. Chiasson, "Why phishing still works: User strategies for combating phishing attacks," IJHC, 2015.

[50] A. Kunz, M. Volkamer, S. Stockhardt, S. Palberg, T. Lottermann, and E. Piegert, "Nophish: evaluation of a web application that teaches people being aware of phishing attacks," Informatik, 2016.

[51] N. A. G. Arachchilage, S. Love, and K. Beznosov, "Phishing threat avoidance behaviour: An empirical investigation," Comp. Human Behavior, 2016.

[52] A. Xiong, R. W. Proctor, W. Yang, and N. Li, "Is domain highlighting actually helpful in identifying phishing web pages?" Hum. Fact., 2017.

[53] M. M. Moreno-Fernández, F. Blanco, P. Garaizar, and H. Matute, "Fishing for phishers. improving internet users' sensitivity to visual deception cues to prevent electronic fraud," Comp. Human Behav., 2017.

[54] S. Gopavaram, J. Dev, M. Grobler, D. Kim, S. Das, and L. J. Camp, "Cross-national study on phishing resilience," in USEC, 2021.

[55] J.-W. Bullee and M. Junger, "How effective are social engineering interventions? a meta-analysis," Information & Computer Security, 2020.

[56] A. Ferreira and G. Lenzini, "An analysis of social engineering principles in effective phishing," in IEEE STAST Workshop, 2015.

[57] D. Ruby, "40+ net ix statistics 2023." https://www.demandsage.com/net ix-subscribers/, 2023, accessed March 23, 2023.

[58] M. Bozyczko, "Amazon in europe: key statistics," https://nethansa.com/blog/amazon-in-europe-key-statistics, 2023, accessed March 23, 2023.

[59] Zalando, "Zalando grows customer base and progresses on platform transition," https://corporate.zalando.com/en/ nancials/zalando-full-year-22-results, 2023.

[60] AirBnB, "The eu host action plan," https://news.airbnb.com/wp-content/uploads/sites/4/2021/12/The-EU-Host-Action-Plan_Dec-2021.pdf, 2023.

[61] FinancesOnline, "Number of active gmail users: Statistics, demographics, usage." https:// nancesonline.com/number-of-active-gmail-users/, 2023.

[62] M. Mohsin, "10 google search statistics you need to know in 2023," https://www.oberlo.com/blog/google-search-statistics, 2023, accessed March 23, 2023.

[63] M. Iqbal, "Instagram revenue and usage statistics," https://www.businessofapps.com/data/instagram-statistics/, 2023.

[64] Statista, "Facebook monthly active users in europe as of 1st quarter 2023." https://www.statista.com/statistics/745400/facebook-europe-mau-by-quarter/, 2023.

[65] LinkedIn, "Statistics—check out the numbers to understand the world's largest professional network." https://news.linkedin.com/about-us, 2023.

[66] PayPal, "Paypal announces nearly 35 million accounts in europe," https://newsroom.paypal-corp.com/2007-03-20-PayPal-Announces-Nearly-35-Million-Accounts-in-Europe, 2023, accessed June 1, 2023.

[67] Mappr, "Countries with uber 2023," https://www.mappr.co/thematic-maps/countries-with-uber/, 2023, accessed June 1, 2023.

[68] GitNux, "The most surprising yahoo statistics and trends in 2023," https://blog.gitnux.com/yahoo-statistics/, 2023, accessed June 1, 2023.

[69] DataReportal, "Twitter users, stats, data and trends." https://datareportal.com/essential-twitter-stats, 2023, accessed July 1st, 2023.

[70] J. Horton, R. Macve, and G. Struyven, "Qualitative research: experiences in using semi-structured interviews," The real life guide to accounting research. Elsevier, 2004.

[71] B. Krishnamurthy and C. E. Wills, "On the leakage of personally identi able information via online social networks," in WOSN, 2009.

[72] E. A. for Fundamental Rights, "Child participation in research," https://fra.europa.eu/en/publication/2019/child-participation-research, 2014.

[73] B. A. Alahmadi, L. Axon, and I. Martinovic, "99% false positives: A qualitative study of SOC analysts' perspectives on security alarms," in Proc. USENIX Security Symp., 2022.

[74] M. Delacre, D. Lakens, and C. Leys, "Why psychologists should by default use welch's t-test instead of student's t-test," IJSP, 2017.

[75] P. López-Aguilar, C. Patsakis, and A. Solanas, "The role of extraversion in phishing victimisation: A systematic literature review," in APWG Symposium on Electronic Crime Research (eCrime), 2022.

[76] D. R. Thomas, "A general inductive approach for analyzing qualitative evaluation data," American journal of evaluation, vol. 27, no. 2, pp. 237–246, 2006.

[77] K. Sharma, X. Zhan, F. F.-H. Nah, K. Siau, and M. X. Cheng, "Impact of digital nudging on information security behavior: an experimental