IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

How Johnny Experiences Phishing Warnings: A Qualitative Study Investigating the Impact of Design Decisions on the User

STEFANIE PHAM[1,2], Gabriele Lenzini[2], and Daniela Pöhn[3]
[1]Ludwig-Maximilians-Universität München, Munich, Germany
[2]Université du Luxembourg (UL/SnT), Luxembourg, Luxembourg
[3]University of the Bundeswehr Munich, RI CODE, Neubiberg, Germany

Corresponding author: Daniela Pöhn (e-mail: daniela.poehn@unibw.de; orcid: 0000-0002-6373-3637).

ABSTRACT

To this day, phishing remains one of the most critical and elusive threats in cybersecurity. Although detection technologies have evolved and improved, they have not kept pace with novel phishing strategies. Thus, when software cannot definitively identify phishing, the last line of defense rests with the user when they are asked to ''think before you click''. The appeal is commonly accompanied by warning messages, supposedly providing the user with enough information and incentive to make an informed, secure decision. However, warning messages must be carefully crafted because their elements can considerably affect the user's agency, trust, and decision-making. We selected four of the key design elements in warning messages: *content*, *placement*, *level of friction*, and *timing*. We conducted a qualitative study using think-aloud sessions with 18 participants. Each participant was presented with phishing scenarios, accompanied by warning messages that differ in regard to those four elements of design, followed by a post-session interview. Thematic analysis revealed 13 themes across the four elements and from the analysis, novel insights emerged. For instance, *timing* changes the context in which users frame their concern: rather than being concerned about the potential consequences of clicking —as the warning intends— they become suspicious of the app displaying the message, fearing it may invade their privacy and violate their security. Our findings form a basis for future research about how to design and implement mechanisms, such as warning apps, that are more adaptable, targeted, and potentially more effective in protecting users from phishing attacks.

INDEX TERMS Phishing, phishing warning, security warning, social engineering, user-centric interaction design, user study.

I. INTRODUCTION

PHISHING is a well-known threat used to perpetrate multi-stage cybercrime against organizations [1]. It is the root cause of financial losses, identity theft, ransomware and malware infections, and reputational damage.

Despite years of research on how to mitigate it, it remains one of the top cyber threats: In 2023, ''the worst year for phishing on record'', there have been almost 5 million phishing attacks, as reported by the Anti-Phishing Working Group (APWG) [2]. This trend has only apparently stopped in 2024, where instead the number of unique email campaigns (*i.e.*, classes of phishing emails with different subjects) was up 64%, proving that ''phishers are diversifying their email subject lines in order to bypass email filtering'' [3].

Since phishing is a semantic attack that combines technical subterfuge with deception (see Schneier [4]), various prevention mechanisms exist to mitigate risks, including approaches like awareness campaigns, as well as technical solutions, such as automatic detection systems. However, no algorithm can provide a categorization with one hundred percent confidence, as shown with PhiUSIIL [5]. There will always be messages that might be suspicious but cannot be confirmed to be fraudulent. In those cases, Johnny, our archetypal end user in security research, is the last line of defense. Ideally, it is Johnny that assesses the legitimacy of the suspicious message, and it is Johnny that decides how to act.

User-based warnings are intended for these edge cases, where automatic detection cannot provide a definitive decision. They are not meant to appear regularly, but to complement automated detection in situations of uncertainty. Thus,

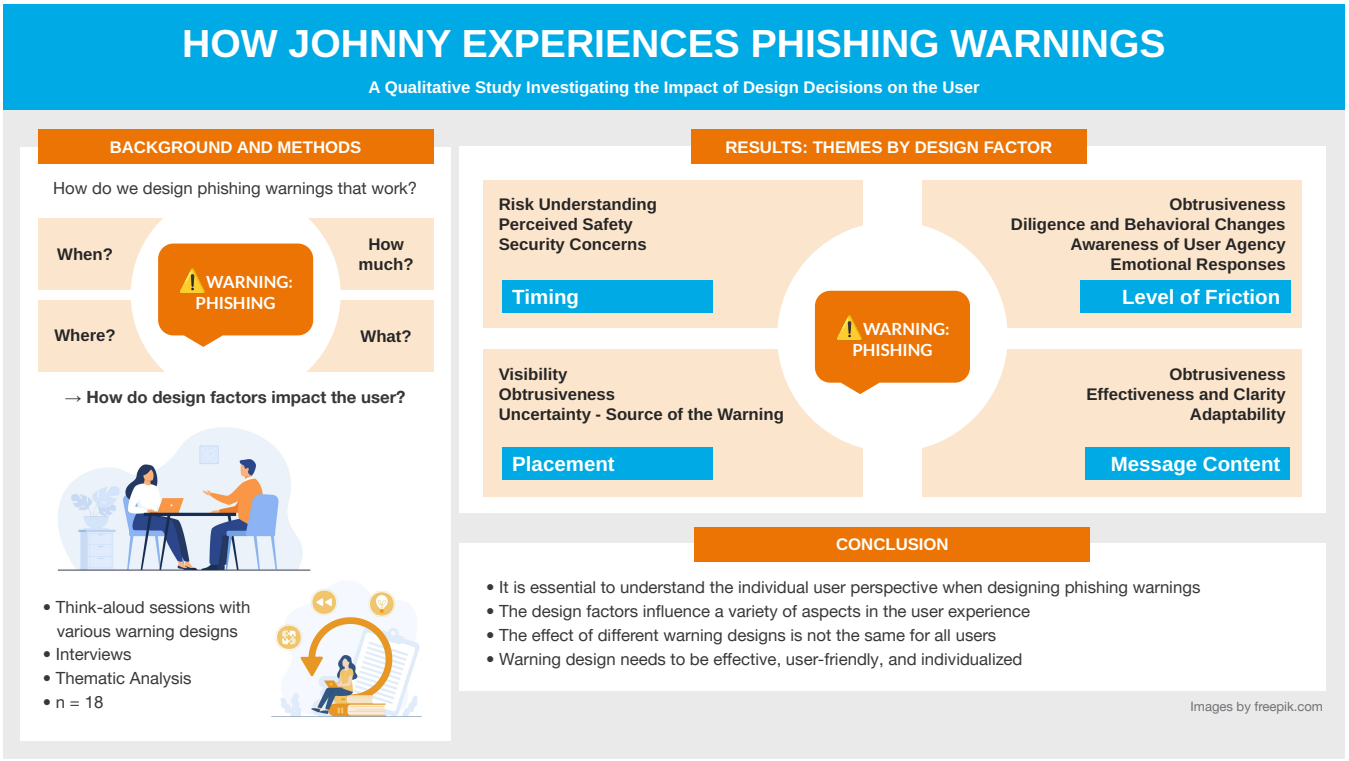


FIGURE 1: We explore four key elements in warning design and implementation: timing, placement, level of friction, and message content. We identify 13 themes emerging across the four factors based on a qualitative study with 18 participants. These help to improve phishing warning design in the future.

automatic detection and user-based warnings should be understood as complementary approaches rather than competing ones.

So firstly, Johnny must be made aware that something ''fishy'' may be happening: in such cases, a commonly adopted solution by browsers, email clients, and similar message-delivering apps is to alert Johnny with a live security warning. The fact is, security warnings are not all the same and do not always trigger the response in Johnny that the warning designer had in mind. Security warnings can irritate Johnny by obstructing his path to the primary goal. Badly designed security mechanisms cannot only fail to engage, but may reinforce Johnny's habit of ''clicking without thinking'' [6].

The difficulty of designing engaging, user-friendly, and effective phishing warnings is a well-known problem. Researchers have proposed a plethora of anti-phishing solutions and evaluated their effectiveness in end-user studies [7]–[9]. Researchers have also measured how effective warnings are at helping against phishing [10], [11], and compared their effectiveness with other solutions [12]–[14].

These types of studies are often *quantitative* and fail to investigate how the user perceives the warnings (e.g., are they helpful or useless?), how they react in response to them (e.g., do they consider them, do they consider them first but then ignore them, or do they ignore them right away?), and why they react that way (e.g., do they believe they are a waste of time or do they take them seriously?).

Das et al. [15] criticize that only 14% of recently published papers about phishing include user studies and that the majority of those studies are on tools developed by the researchers in support of novel ways to cope with the threat, but without a qualitative validation. Studies that aim to understand how specific design factors impact the user's perception and response to these warnings are rarer, and it is in this particular research field that this paper contributes to. We think that this understanding is essential in order to be able to design more effective warning-based anti-phishing solutions.

We focus our research on four specific elements that can impact a person's perception of and reaction to a warning (see, Section 1, and Section II) (i) the warning content and adaptability [16], [17]; (ii) the placement of the warning [18]; (iii) the level of friction induced by the warning [19]; and (iv) the timing at which the warning appears [20].

The influence of these four elements has not been studied in the context of phishing so far, so when contextualizing our contribution over the four factors impacting phishing warnings, a major *research question* arises: How do message content, placement, level of friction, and timing impact the user's perception of the risk and the user experience of a warning, and how do they influence the following actions? This work contributes to answering these questions.

There are previous considerations to take into account. *Message content* does not seem to improve user self-

protection [16], but also because of lack of adaptability, which is indeed praised as noteworthy [17]. Thus, we study adaptability with message content in phishing. *Placing* a warning closer to where the threat may materialize seems to improve the effectiveness of the message [18]. We study how Johnny's perception changes depending on where the warning will appear: close to where he is going to click or type, or somewhere else, e.g., on the taskbar. *Friction* has been introduced as a concept by Distler et al. [19]. We will study the effect of different friction levels, specifically in phishing warnings. *Timing* is defined by when a warning will be shown. Warnings that appear closer to a dangerous link are more effective [18] and this effect has been studied mainly for online gaming security, see Qu et al. [20]. To our knowledge, we are the first to study its effect on phishing warnings. Moreover, we look into the perception of privacy with timing, because a better timing may suggest that more information about the user has been processed.

Studying the four mentioned elements by using a methodology explained in Section III, we *contribute* to advance the state of the art, as follows: (Section IV):

(i) we propose a first analysis of the influence of the variables of timing, placement, level of friction, and message content on phishing warnings;
(ii) we identify 13 themes across the four variables based on qualitative think-aloud sessions and a thematic analysis;
(iii) we discuss valuable insights for phishing warning designs: the variables of timing and context can lead to security concerns, however, no universal solution to phishing warning design is possible.

Although this work is a scientific contribution to the field of usable security, it provides practical information for designing real-world phishing prevention mechanisms. Thus, by extension, this work contributes to better protecting users from phishing attacks (Section V, and Section VI).

II. BACKGROUND AND RELATED WORK

We discuss in more detail related research on mitigation strategies for phishing and the four elements of design that this paper studies: timing, placement, level of friction, and message content (including its adaptability).

A. MITIGATING PHISHING THREATS

Among the multiple approach and strategies to mitigate phishing threats that exist (see Naqvi et al. [21]) two main categories of strategies stand out: non-technical and technical.

A widespread non-technical strategy, largely adopted within organizations, is education. Employees are trained to respond to phishing campaigns by taking part in simulated phishing exercises (see Siadati et al. [22] and Kumaraguru et al. [23]) where their resilience is put to a test. Educational modules are offered to those who fail the test, that is, that fall for the phishing, usually by clicking on a link.[1]

[1]To wait until they edit a password on a form would be controversial, even illegal if performed without consent, according to the GDPR.

Whether training is an effective solution is still unclear, the subject is quite heated, but the quest for effective training sparked research on different training methods (see Jampen et al. [24]), or on innovative strategies, such as gamification (see [25]–[28]).

The second cornerstone of phishing prevention is technical. Stand-alone or offered as browser extensions, modern phishing detection algorithms utilize a combination of different strategies, including (without the ambition to be complete) the following.

- *List-based*: A classical strategy that keeps a list of known phishing websites (blocklists) alongside a list of legitimate websites (allowlists). Examples of systems list-based are "Google Safe Browsing" and "PhishTank" [29].
- *Feature-based*: It analyzes specific attributes of a website or email to identify phishing. Quite accurate in general, but not very flexible, an example of a system using this strategy is to search for abnormalities by comparing the address of resource elements on a web page with the URL, as shown by Moghimi and Varjani [10].
- *Machine learning-based*: Here, machine learning models are trained on phishing attack data to be able to predict whether a website (see Srinivaso Rao and Pais [30]) or email (see Bountakas et al. [31]) is fraudulent. Unlike feature-based systems that work with predefined conditions, machine learning systems are more adaptable, and can learn.

Although anti-phishing technologies are getting better (*i.e.*, have higher detection capabilities and lower false positive rate), there will always be some phishing message that escapes, that is the true positive rate remains below 100% (Gmail's spam filter has 99.9% accuracy according to Dada et al. [32]). This is why warnings are still elementary in phishing: in that gap where humans have the responsibility to detect a phishing attack, warnings can pass along valuable information and can act as a catalyzer in raising a user's defenses. The modality in which this indicator is designed is as important as the presence of the warning itself.

B. SECURITY WARNINGS DESIGN

The design of warnings that are effective is a complex task. We recall the most relevant research on the subject, within the scope of our work.

Security warnings, like those that browsers show sometimes, can be effective, but only if users have some prior security experience [33]. Johnny's experience (e.g., being a novice or an expert), and thus his mental model of the situation, determine his behavior in response to a warning [6]; but there are no current warning designs that blatantly violate accepted good design practices [34].

Researchers can tell us indirectly about the effectiveness of a warning design when they validate an anti-phishing solution they propose. Volkamer et al.'s "TORPEDO" [8], plays with proximity and frictions to offer warnings ("tool-tips") that

are hard to ignore, while observing that users improve their detection rate. ''MailTrout'' [7], a browser extension, also reaches high accuracy and user acceptance rates with AI-based warnings whose content explains the threat, describes cues, and recommends actions. Describing what cues justify the warning seems to increase a user's detection and speed of detection rate. In fact, similar improvements are reported by Cooper et al. [9], where the authors study the effectiveness of audio and visual cues in their tool ''PAWS'' (Phishing Alert and Warning). Instead of working with content, Aggarwal et al. [35] work on placement. In their ''PhishAri'', designed to be integrated on X (formerly Twitter), warnings work by displaying a green (resp., a red) circle next to any safe (resp., potentially phishing) link.

C. ROLE OF TIMING, PLACEMENT, FRICTION, AND CONTENT IN SECURITY WARNINGS

Several researchers have studied how the four warning design variables of timing, placement, friction and content influence user perception of security risks and behavior.

a: Timing

Qu et al. [20] demonstrate the importance of timing for the effectiveness of security warnings. The purpose of their warnings was to encourage users to make secure decisions when deciding on a password in an online game. According to their findings, displaying security warnings later in the user's journey is more effective than displaying them earlier. Greco et al. [36] show that the warnings should be displayed after seeing the email content, while Bender et al. [37] emphasize on just-in-time warnings. With a focus on URL restriction, Petelka et al. [38] used time delay as one variable.

b: Placement

Petelka et al. [18] showed the benefits of placing phishing warnings close to the link rather than in a banner at the top of the phishing email. This placement can improve the effectiveness of the warnings by associating the warning message with the actual hazard (*i.e.*, link) within the email.

c: Level of Friction

Cox et al. [39] argument for friction in the form of micro-boundaries, as they intentionally create pauses in the interaction flow that may create more mindful behavior. Slifkin and Neider [40] conclude that interrupts help to classify phishing emails. Distler et al. [19] present a framework for understanding security-enhancing frictions. Bravo-Lillo et al. [41], [42] discuss the concept of pop-up fatigue that describes the diminishing attention of users when repeatedly encountering pop-up notifications. The authors also investigate whether integrating elements that capture attention into the design of dialogues can increase user attention. They found that the attractors that forced user interaction were resilient to habituation. Dennis and Minas [43] explore cognition-related aspects of security decisions, referring to non-conscious automatic cognition (System 1) and deliberate, logical, and conscious thoughts (System 2; Kahneman [44]). The authors challenge the assumption in traditional security research that users are making deliberate decisions by arguing that warning messages require System 2 thinking to process what auditory cues could trigger. This would constitute an extreme form of friction, surpassing the friction levels suggested by Bravo-Lillo et al. [41], [42].

d: Message Content

Harbach et al. [45] investigated the readability of real browser warning messages. The authors identified issues in the comprehensibility of the messages due to technical terminology and overly long or complex messages. The authors underline the importance of clear, short messages, that do not rely on a technical vocabulary. Zaaba and Boon [46], studying usability issues of security warning dialogues, found that the content of warnings, to be effective, must clearly instruct the user on the next safe step. However, as Lain et al. [16] proved by conducting a large-scale, long-term phishing experiment involving more than 14,000 study participants, conciseness also matter because detailed warnings do not significantly improve effectiveness. They study does not provide an explanation for this conclusion. Greco et al. [36] summarize that explanation messages in the warnings should be a included. Sarker et al. [47] propose that ineffectiveness of a message content is affected by explanations that are (a) insufficient and unsatisfactory; (b) lengthy; (c) incomprehensible; and (d) inconsistent. High quality message contents should satisfy the same qualities as good pieces of writing do.

D. DISCUSSION OF THE RELATED WORK

We conclude our review of the state of the art with a brief discussion. The literature reveals that crafting effective warnings is challenging, and suggests that various more nuanced questions must be addressed if we wish to design warnings that are more effective in guiding users into decisions that protect them from phishing. Much of the existing work is quantitative, leaving a notable gap in the qualitative understanding of those questions (see also Sarker et al. [47]).

Prior studies have explored elements that could affect the design of security warnings. We are interested in particular in *timing*, *placement*, *friction*, and *message content*. However, these aspects have partly not been investigated in the context of phishing, as summarized in Table 1. For example, the three studies that investigated timing in the design of phishing warning show that warnings should be shown later, either after seeing the content of the email or just-in-time. The study of friction in security warning is at its beginning, and there remains a need for a more comprehensive understanding of how friction impacts user experience. On the topic of message content, research suggests that explanations are useful, while more detailed warnings do not increase effectiveness without giving reasons for this observation.

This paper's research is novel because it studies these particular design factors in the context of phishing warnings with an in-depth, qualitative analysis of user experiences. This

TABLE 1: Comparison of related work related to the variables of timing, placement, level of friction, and message content.

Study	Variable			
	Timing	Placement	Level of friction	Content
[20], [37], [38]	x			
[36]	x			x
[18]		x		
[19], [39]–[44]			x	
[16], [45]–[47]				x
Our study	x	x	x	x

user-centric approach could significantly aid the development of effective and user-friendly phishing warnings.

III. METHODS

We recall our research design, study materials, participant recruitment, data collection and analysis methods.

A. RESEARCH DESIGN AND PROCEDURE

We opted for a qualitative study on how *timing*, *placement*, *level of friction*, and *message content* influence user experience with phishing warnings. The goal is to gather insights into the feelings and reactions of the participants when the four selected design variables are altered. We conducted a series of in-depth, one-on-one interviews, and we analyzed the data using *thematic analysis* guided by Braun and Clarke [48]. This guide led to the following steps.

1) Familiarize with the data through reading and note-taking.
2) Develop codes that capture ideas from the data relevant to the research questions.
3) Group the codes into potential themes that capture broader patterns in the data.
4) Review and refine (e.g., combine, separate, or discard) those themes to represent coherent patterns in the data.
5) Give them names and definitions, as well as in-depth explanations.
6) Contextualize all the themes in a narrative that analyses the data and is supported by data extracts (quotes). This narrative reflects on the research questions.

1) Research Protocol

Phishing-related studies frequently use deception because it makes it possible to observe genuine reactions. However, we opted for a non-deceptive approach: under the same scenario, it allows us to have a transparent and open discussion with our participants on all aspects of our research. In fact, the research protocol was developed to encourage the participants to verbalize their thought processes. We also prompted specific open-ended questions to the participants to invite them to share their thoughts and reflections related to the research questions.

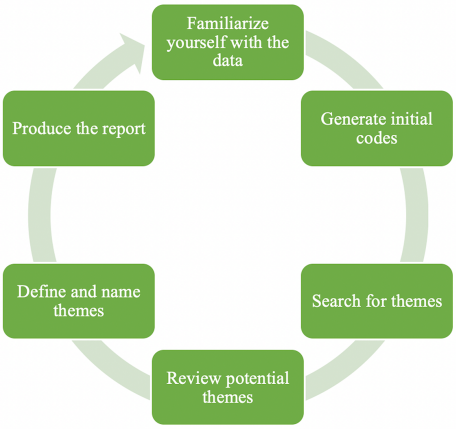


FIGURE 2: The continuous cycle of Thematic Analysis [49].

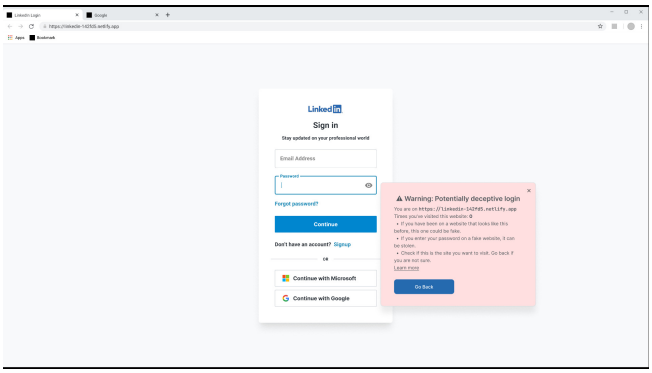


FIGURE 3: Baseline warning, in its full graphical display.

The research protocols began with welcoming the participant to the lab. After we explained the goal of the research, we clarified the definition of phishing, to give the participants a common background in the research context. The participants were then guided to take note of the information sheet, fill out a consent form, and answer a demographic questionnaire.

Then, we allocated time to discuss the topic of phishing. As a warm-up phase, the participants were invited to sit in front of a laptop, telling them that this scenario was meant to reproduce them being at work. Then, they were presented with a series of mock-ups displaying web pages with phishing attempts. The first web page included no warning, followed by a page including a baseline warning (see Fig. 3). Then the participant was presented with several mock-up web pages, each modifying the baseline warning's variables of *timing*, *placement*, *level of friction*, and *message content*. The mock-ups after the baseline warning were shuffled for each participant to avoid order bias. The procedure concluded with explicit questions about the studied variables.

To prevent participant fatigue, we limited each session with the six mock-ups to last at most 45 minutes.

2) Pre-tests

Five pre-tests were conducted that revealed some issues with the comprehensibility of some questions, which were rephrased for the main study.

Stark differences were observed in reactions to message content between tech-savvy and non-tech-savvy participants. This observation led to the introduction of the concept of adaptability to make the warning mechanism suitable for different levels of tech knowledge.

3) Scenario

Some basic context was given to the participants and a simple setup with a classic phishing attack was presented. The purpose of the setup was to allow for the discussion of the warning design factors. An office scenario was selected, which is unrelated to personal online behavior, allowing for more comparability. Participants provided informed consent and were aware that the study examined phishing warnings, the study design was not deceptive.

4) Warnings

The warnings are designed as if they originate from a pre-installed browser extension. They are meant to be displayed when the phishing detection technology behind the extension is not able to decide whether a link and the website behind it are safe or unsafe with certainty (see Krol et al. [50]). The decision-making agency is then relayed to the user.

Mock-ups of a website and browser extensions were designed using the interface design tool [51] as study material. Each variable was visualized in an example scenario. These example pages were used to discuss how the participants experienced changes in these variables. The website chosen for the scenario was LinkedIn [52] as it was the most common social media network to be used in phishing attacks, according to the Check Point Research Team [53]. The mocked phishing pages were designed to be believable but contained minor errors, including spelling mistakes.

The first mock-up (see Fig. 4a) displays a phishing page without any warning. This mock-up was used to establish an understanding of the user's reaction to a fraudulent login page and interpretation of the situation and cues without any assistance.

Fig. 4b shows the same website with a warning from a browser extension used as the *baseline*. The classic design follows the guidelines by Black et al. [54]. The headline ''Warning: Potentially deceptive login'' uses the phrasing ''deceptive'' instead of ''phishing'', modeled after the message Mozilla Firefox shows when blocking a phishing website [55]. The word ''potential'' was added to indicate the uncertainty. The warning highlighted some information to aid the user in making an informed decision. The URL was shown in bold (see [56]–[58]). Similar to the ZecOps Research Team [57], the following line informed the viewer about the number of times they have visited the website. A ''Learn more'' button expands the warning to include concrete reasons why the website is suspected to be phishing by the tool (see [57], [59]).

The mock-up in Fig. 5a was designed to investigate the impact of *timing*. The warning occurred a step earlier in the phishing attack lifecycle, namely in stage 2 (Phishing) instead of stage 3 (Infiltration) in the baseline warning. The user sees an email inbox with a phishing email that includes a fraudulent link and is presented with a warning. Fig. 5b presents a phishing page with the modified *placement* of the warning. The message window is located in the top right corner of the browser, where phishing warnings from browser extensions are commonly placed, e.g., by Netcraft [56].

The mock-up in Fig. 6a depicts a warning with an increased *level of friction*. A micro boundary was introduced, the login was disabled, and an additional step of interaction with the warning (click a button) was needed to proceed.

The mock-up in Fig. 6b features a warning with a modified, shorter *message*. Compared to the baseline message, this warning does not inform the user how many times they have visited the current website. Furthermore, the shortened message provides a concise warning without detailed explanations. This shorter warning assumes the user is already knowledgeable and only requires a reminder. Additionally, the warning message is *adaptable*, for example, based on their level of technical knowledge.

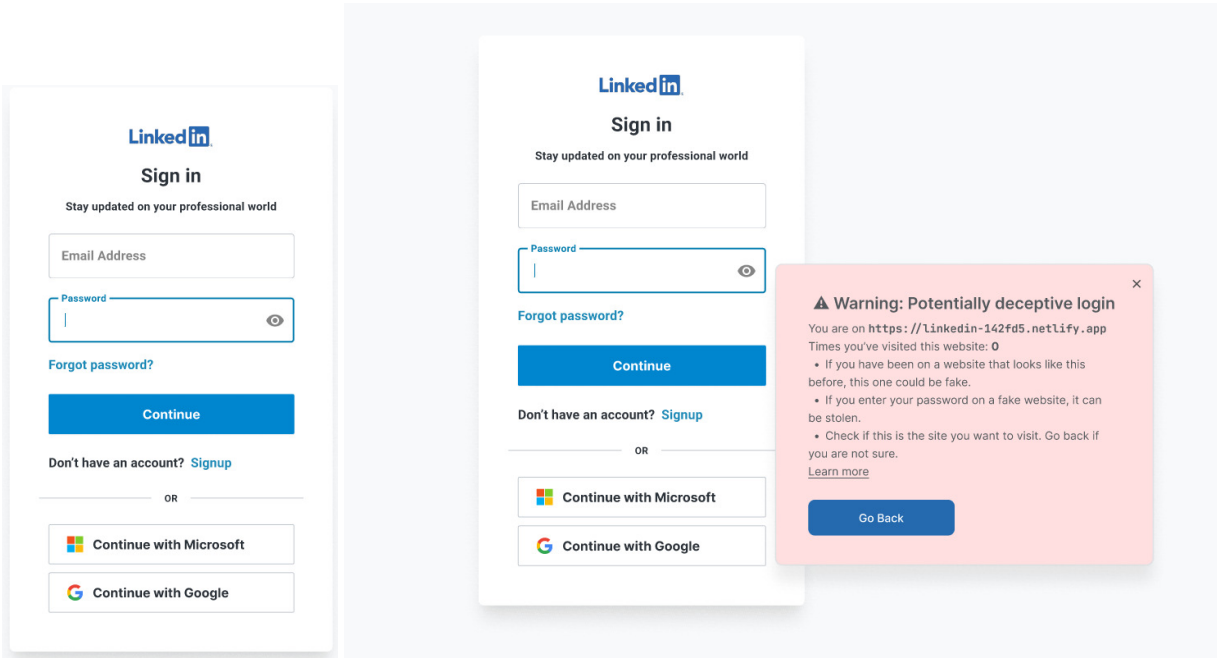
The mock-ups do not present suggested warning implementations; they have been created to illustrate design factors for discussion with the participants.

B. PARTICIPANTS

The in-person study was conducted at the University of Luxembourg's campus in Belval. The participants were recruited through flyers distributed through university platforms to students and staff and a local sports club. They were required to be over 18 years of age, fluent in English, and have an occupation that includes reading emails. As compensation, they received vouchers for online shopping. From the interested applicants, a group with diversity regarding employment status, gender, and technological background was selected. Due to the described recruitment method, the sample may have been prone to self-selection bias.

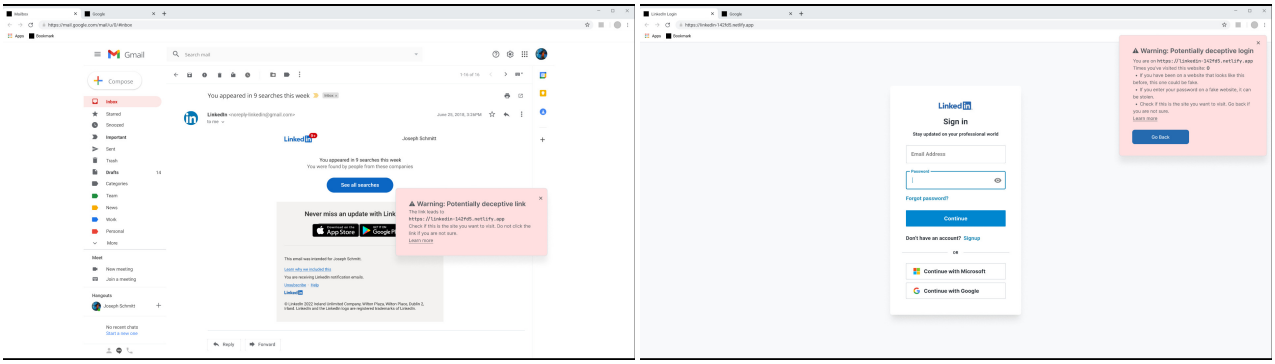
Our sample size of 18 is consistent with recommendations in the literature, which indicate that thematic saturation in qualitative interviews can often be achieved with 9–17 participants [60].

Pseudonymized demographic data of the selected 18 participants is provided in Table 2. The gender distribution was slightly skewed towards males (seven females (39%), eleven males (61%), and no other participants). The ages of the participants ranged from 22 to 55 years, with a median age of 30 years. The sample was slightly skewed towards individuals who reported having technology-related backgrounds (61%), with one participant in the cybersecurity-related field.



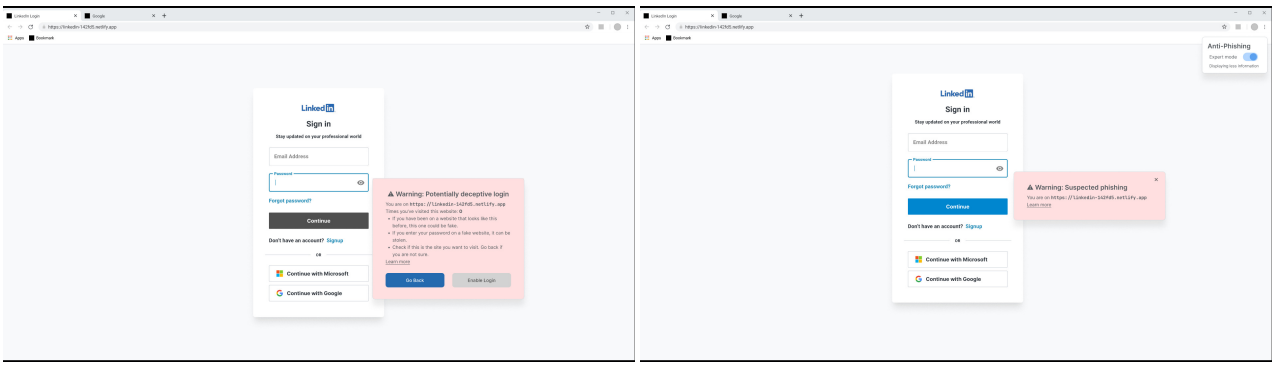
(a) No warning. (b) Baseline warning.

FIGURE 4: Detail (no desktop view) of the no warnings and baseline design prototypes.



(a) Prototype warning with modified timing. (b) Prototype warning with modified placement.

FIGURE 5: Prototype warnings with different variables.



(a) Prototype warning with modified level of friction. (b) Prototype warning with modified message content.

FIGURE 6: Prototype warnings with different variables.

TABLE 2: Interview participant demographic data.

No.	Gen.	Age	Job Title	Field	Education	Tech field	Security field
P1	F	22	Working student	Psychology	Highschool	No	No
P2	M	23	PhD candidate	Computer science	Master's	Yes	No
P3	F	22	Working student	History	Highschool	No	No
P4	M	27	Working student	Engineering	Bachelor's	Yes	No
P5	M	24	Working student	Computer science	Bachelor's	Yes	No
P6	F	29	PhD candidate	Biology	Master's	No	No
P7	F	27	Working student	Computer science	Highschool	Yes	No
P8	M	29	PhD candidate	Computer science	Master's	Yes	No
P9	F	30	Attorney	Law	Master's	No	No
P10	F	31	PhD candidate	Cryptography	Master's	Yes	Yes
P11	M	30	Software engineer	IT	Master's	Yes	No
P12	M	42	E-learning specialist	Education	Master's	Yes	No
P13	M	29	Web developer	IT	Bachelor's	Yes	No
P14	M	38	Tech transfer officer	Intellectual property	Master's	Yes	No
P15	M	30	Project manager	IT	Master's	Yes	No
P16	M	33	Project communications officer	Advocacy	Master's	Yes	No
P17	F	38	Team assistant	Administration	Bachelor's	No	No
P18	M	55	Safety manager	Aviation	Master's	No	No

C. DATA COLLECTION AND ANALYSIS

After having collected the participants' written consent and informed them about the nature of the research, we proceeded with the interviews of the 18 participants in January 2023. Each interview lasted approximately 45 minutes. The interviews were audio and screen recorded, transcribed, and manually refined. Data were pseudonymized in accordance with the GDPR recommendations.

We analyzed the data by following a thematic analysis [48]. This is a common choice whenever the goal is to find out from qualitative data people's opinions and experiences. To evaluate the interviews, thematic analysis was applied to the data with the help of MaxQDA. The transcripts were first segmented according to the four design variables (timing, placement, level of friction, and message content) and impact factor. The initial impact factors are the user's understanding of the warning and the perceived risk, their following action, and their user experience. The additional category of *emotional response* emerged in the coding process. Following the segmentation, themes were developed from patterns across the codes and segments to capture considerable patterns. The codes and themes were reevaluated based on their thematic relevance and frequency across the data.

D. LIMITATIONS

The qualitative study used interviews and mock-ups to simulate an office environment. The responses might be different in an actual work environment. The participants were recruited on the campus in Luxembourg using flyers, which might include a self-selection bias. Also, the number of participants with technical backgrounds is higher. Lastly, the responses might differ from those in other countries.

IV. RESULTS

Here, we showcase our findings, discussing the themes emerged from the thematic analysis of the collected interviews. The results are structured by factor, with three themes having emerged in connection with the factors *timing, placement*, and *message content*, and four in connection with the factor *level of friction*.

A. TIMING

1) Risk Understanding (T1)

Description: Users have a more concrete mental model of the threat at a later stage in the phishing attack, which leads to more compliance with late-stage warnings.

The interviews suggest that presenting the warning later in the user's interaction process with the phishing attack may lead to more concrete comprehension of potential consequences and an elevated perception of risk. Certain participants who dismissed the warning in their email inbox were heeding it at the login web page. Their mental model of the threat appeared to be more concise at a later stage than at an earlier stage. When presented with the warning earlier, in the mailbox, some participants did not perceive the action of following the link as risky enough to be deterred, despite the warning. One participant illustrated this indifference by stating, *"It's not really important. If I will get - okay, maybe you will lose your credentials. Maybe you install something. I will follow, I will just click on it and I will check if it's LinkedIn or if it's something else"* (Participant 4). Participant 1 describes her comprehension of the situation with more uncertainty: *"Um, the warning tells me that there's a deceptive link. But if this is the email, then I don't really see the link that would be deceptive. Okay. Um, but I guess, I mean, if you press... but it doesn't seem that obvious to me how this could be a scam"*. When the interviewer asked her to assess the potential consequences of following the link, she confirmed that she did not associate any risk with the link itself, stating *"Um, I don't know. I think if you would now click on the link, I wouldn't know what would happen or what they would have access to just by doing that because they already have your email..."* She followed up with a vague description of potential consequences and a behavioral remark, *"I feel like*

FIGURE 7: Overview of the themes.

maybe I would just click 'close' on that warning".

Conversely, when the warning was shown at the login stage in the baseline warning mock-up, both participants were aware that logging in on the website was unsafe and that they would not proceed to enter their credentials. When asked about their comprehension of the login stage warning, they commented *"It makes me think that something is wrong... and I will not sign in here"* (Participant 4) and *"It's trying to tell me that ... it could be a fake scam web page trying to get me to enter my email address and password for an account I've already created on a different website"* (Participant 1).

2) Perceived Safety (T2)

Description: Early-stage warnings improve user experience for users who already perceive following malicious links as unsafe because the early warning makes them feel safer.

Similarly to the previous theme, this one also stems from the understanding that the security risks are higher at a later stage in the phishing attack. However, contrary to the previous theme, this theme deals with participants analyzing the risk and coming to the conclusion that earlier warnings feel safer to them and therefore improve their user experience when compared to later-stage warnings. The concrete hazards of following a phishing link without entering your credentials are elaborated upon in Section V-A.

Participants described feeling like they were at a lower risk when presented with the warning in their email inbox rather than at the login page. The degree to which they were aware of concrete risks from just following the link varied. This perception seemed to stem from a desire to be shielded from potential threats before interacting with a potentially malicious website. Participant 6 was not entirely certain about the possible consequences, but expressed that she preferred to not even interact with a malicious website: *"For me, [the warning] is preferable in the email itself because I don't want to accidentally open up something. I'm not sure if they can obtain information ... just based on that."* This highlights an underlying uncertainty about the threat which, for this interviewee, led to caution.

Other participants described concrete consequences associated with interacting with phishing emails, even before providing credentials. Participant 2 pointed out his concern with information other than his credentials being stolen, stating, *"...Even if somebody clicks on it, and he may close the browser, still some of his information gets stolen from the - from his network, like his IP address and stuff."* Similarly, Participant 15 feared that he might be tracked: *"Because I know if I click, then they're going to track me. They know I have come from the email."* Consequently, Participant 2 also preferred to be warned at an earlier stage, like Participant 6. Participant 15 preferred not to be warned at all, stating he would be able to detect the phishing attempt without assistance and exit the situation at an early stage as well.

3) Security Concerns (T3)

Description: Displaying a warning in an email environment triggers security and privacy concerns associated with browser extensions.

When the warning appeared in the mail client rather than on the login page, participants started expressing security-related concerns about the browser extension. Presenting the warning in the email inbox seemed to draw attention to permission issues because emails are perceived as more private. With the warning appearing in the inbox, participants began pointing out the fact that the extension had access to their emails. Participant 15 articulated this concern, mentioning, *"My first thought is it's cool. But my second thought is, now that it can read my emails, ... is this thing reading something else? Which would encourage me to delete it."* This sentiment was

echoed by Participant 16, who was uncomfortable with the extension's access to personal emails but raised an additional concern beyond the extension itself. His fear was that even if the extension was not invading his privacy, it may introduce a vulnerability into his system: *"If you have an add-on installed on your browser that is not updated on a regular basis, it may lead to backdoors that other people can exploit and then gain access once again to my personal information."*

There were also participants who, after pointing out their security concerns with the extension's permissions, seemed to accept them. Participant 3 reflects, *"I don't know how I feel about the browser being able to access my emails in real time ... but if it's not done to steal my data and sell it to a third party company, I think I'm fine with it ... I don't know if I would care that much since we disclose already a lot on social medias and whatnot."* When presented with the timing mock-up, Participant 15 also expressed his desire to have control over what extensions or add-ons were installed on his devices. He remarked: *"If it's pre-installed, then I would expect to know what this is... I would like to know what's installed on my computer and who is an admin of my device, who can install things, who can see things on my computer and stuff like that."*

B. PLACEMENT

1) Visibility (P1)

Description: The warning's placement determines its visibility and the perception of its relevance. It is more noticeable when placed closer to the login field. In the upper right corner of the browser, it can be dismissed not only because it is less noticeable visually, but also because the position is associated with other content.

The warning's placement influences the visibility and triggers associations with other content in the participants. Multiple participants pointed out the decreased and increased visibility of warnings placed in the top right corner and near the login field, respectively. They described the latter placement as *"more visually striking"* (Participant 17) and that *"your eyes are primarily focused on the email address and password fields"* (Participant 18). Participant 6 expressed a personal preference for this placement for that reason: *"For something like this, I would actually prefer it probably to be closer to where I'm anyway focusing my attention. So otherwise, I'm more likely to miss it."* Participant 17 added to that perspective that she might disregard the warning in the corner, especially in stressful situations or when she is in a hurry.

Some statements about the disregard of the corner warning were not directly related to the lack of visibility, but rather to their experience of content in that position in their browser. They explained that they might mistake the warning for another type of notification, such as an advertisement or a browser update reminder (Participants 13 and 17). Participant 8 associates this placement with browser extension notifications, saying: *"This has become like a white noise zone, so I would ignore it."*

2) Obtrusiveness (P2)

Description: The warning's placement can negatively affect the user's experience if it clutters the screen; it can also distract from crucial information when it is placed prominently.

Some participants perceived a warning placed close to the dangerous elements of the phishing page as more disruptive and distracting, and a more subtly positioned warning as less disruptive. One explained, *"It feels ... less intrusive ... and it looks more standard. What I mean is, it's more common to have ... these kind of warnings off the main action you're trying to perform. So it feels less annoying for some reason. Maybe it's because I'm used to it"* (Participant 12). The familiar placement appeared to be more comfortable. In contrast, the warning placed close to the login field was perceived as cluttering by some users. *"I think it's a good way to be informed, but there's a lot of information on the screen ... I don't know if having a big pop-up on your screen is an effective way to be informed"* (Participant 17). Another participant did not share this sentiment and preferred the placement because the warning *"basically didn't cover any real information, but still intruded on the main window"* (Participant 18). On the login page where it appeared, the warning did not cover anything other than white space due to the design of the web page.

A new consequence of the warning's placement emerged when Participant 15 was presented with the warning in the email inbox, which was positioned close to the phishing link in the mocked deceptive email. The prominent positioning of the warning appeared to have distracted him from noticing the sender address, which was a notable phishing cue in the example email. It occurred to him later in the discussion: *"Something I haven't actually picked up on is the fact that the mail is - in the first place, the mail is wrong. And what is this, 'gmail.com'? 'noreply.linkedin@gmail.com'? What is this? Okay, why are we even talking about this? Yeah, I missed that somehow because I was looking at other places. I was looking at the link. I haven't paid attention to this, even though this is where I should have looked at..."* This realization caused an emotional response, indicating distress over overlooking this critical detail.

3) Uncertainty about the Source of the Warning (P3)

Description: The warning's placement on a web page can give users clarity about its source, or arouse suspicion if it does not.

Users utilize the warning's placement to distinguish between browser extension warnings and website-generated content. They are able to identify the warnings as an anti-phishing browser extension when the warning is placed in the top right corner, overlaying the browser UI, but have misconceptions or raise doubts about the warning's origin when it is placed on the web page close to the login field or the link.

One participant had the misconception that the warning was coming from the website itself, replying to a question about his understanding of the warning with *"to be very honest, I don't know. But I think it shows that LinkedIn is taking care more about security to their users [sic] ... So*

if LinkedIn provides this type of warning, I will say I feel safe" (Participant 4). He assumed he was on the legitimate LinkedIn website as he had not detected the phishing cues, and gathered the warning originated from LinkedIn itself. Other participants actively questioned the warning's source. The thought occurred to Participant 15 after being presented with a warning close to the login field: *"The more I look at this, the more we talk about this, the more think about this, I would think maybe there's something weird going on."* He then became suspicious of the warning, assuming it to be fraudulent itself, explaining, *"That's a meta-inception-phishing kind of thing, because ... this is the window of the page ... If it has something to do with the browser itself, it would pop up as a notification from the toolbar."*

Another participant echoed this sentiment immediately when confronted with a warning close to the login field. He voiced his concern: *"Yeah, it's very visible... But my question is, where does it come from? If this is just an extension over the browser? If I'm seeing it for the first time, I would be like, okay, that's another, that's a part of the scam for me"* (Participant 14). He immediately reached the conclusion that the warning itself was fraudulent.

When presented with a warning in the top right corner of the web page, overlapping the browser UI, Participant 6 stated his preference for this design: *"Yeah. So the position of the warning definitely, it seems not to be coming from the web page because it covers the part ... of the browser UI. I'm very happy with that."* All participants who raised this concern confirmed that this placement would ease their suspicions in that regard.

C. LEVEL OF FRICTION

1) Obtrusiveness (F1)

Description: The level of friction of the warning influences how obtrusive it feels to the user. Introducing a microboundary can be perceived as overbearing or unnecessary, particularly if they are shown on platforms that are not considered sensitive.

The increased friction was sometimes perceived as too obtrusive or unnecessary. Some participants expressed that the microboundary felt like a disturbance and that it was an unwelcome addition for their user experience. Participant 16 stated, *"If I had the ability to turn that specific part of the ...extension off, I would. If I couldn't turn it off, and it was there by default, it would be slightly annoying."* He went on to explain that the microboundary was not necessary for him to make safe decisions, *"I feel that I am not too likely to ignore the popup itself. I would still check, for example, the URL. So I don't feel that there is added value for me as a user."* For Participant 4, the necessity of friction was dependent on the platform's significance: *"It should not be on every website. It should be on some sensitive websites ... developers always think that they are providing a more user-friendly, safer service. But if it's not very important data, the user can sometimes get irritated by this warning."* He elaborates on how he evaluates the relevance of the threatened data for the scenario example: *"If you are a well-established professional, it might not be very important for you... But if you're newer in your career, like me, with one to three years of experience, it's really important."* His classification was solely based on the immediate access to his account. This illustrates how a low sense of perceived risk lowers the acceptance of obtrusiveness.

2) Diligence and Behavioral Changes (F2)

Description: The level of friction influences the user's diligence and their considerations for their next action. Increased friction triggers users to double-check warnings.

The change in the warning's level of friction led to users engaging with the warning more consciously. Participant 8 describes this heightened awareness after clicking the 'Enable Login' button: *"Now I'm physically doing something to acknowledge the risks ... this will stay more in my mind as opposed to just making a mental note."* The extra step in the process made the warning harder to dismiss or ignore on autopilot-behavior. The increased friction might lead to a switch from System 1 to System 2 thinking and compel users to actually read the warning, with Participant 7 saying, *"I mean, if this one actually stops me from entering information, there is a higher chance that I will actually read what's written there."* Participant 2 articulates how the friction would lead him to evaluate his next step more carefully: *"People will at least think twice about doing this ... because at first there was only - you can just write and continue and just ignore this, but now ... they will definitely check again before doing 'Enable Login', like a two-step verification type of thing."* With their heightened diligence, some participants were now considering double-checking their decision externally: *"My instinct would be to contact the IT department"* (Participant 17).

3) Awareness of User Agency (F3)

Description: Introducing a microboundary increases the user's awareness of their own agency in the phishing scenario. It makes users more aware of the fact that they have the choice to proceed or to go back to the web page.

The introduction of a microboundary raised the participants' awareness of their own decision on how to proceed. With the increased friction, people started addressing the user choice between going back or interacting with the warning to proceed. This was despite the fact that all mocked warnings offered the user a choice, this one simply required an additional action to choose to proceed. The increased friction seemed to draw focus to the fact that users had agency in the scenario, and made some doubt that they should. Participant 8 articulated, *"So yeah, it gives me a little bit more options ... so I would prefer to have other options other than 'Go Back', like 'Enable Login'. So I understand [the warning], and now please allow me to just do my stuff if I want to trust and go ahead."* When asked, the participant confirmed he understood the mechanisms of both options with different levels of friction, and continued: *"Yeah, this is now putting*

the options forth explicitly. Before it was a little bit of an ambiguous implicit decision... I have clear options. Either I enable the login or go back." At first, Participant 10 stated she preferred the design with the increased level of friction, naming the user agency as a reason: *"So it gives you the warning, and then it's your duty to make sure that this is the right website or this is the wrong website. So it's your choice ... so I think in general it's good to have this option so that it's your choice what you want to do."*

However, when discussing the scenario more, she started raising concerns about the user agency in the context of the scenario: *"So I think in general it's good to have this option so that it's your choice what you want to do. But in this scenario for LinkedIn, I think it's better not to have the choice to sign it. I don't know, as I said, because of the address, because I know that this is not the LinkedIn address. So I expect the warning to somehow don't give me the choice."* Participant 17 shares this opinion, circling back to the professional environment the scenario took place in. She underlines why leaving security decisions to individual employees is risky: *"I think this could be beneficial, but in a professional context, it would be crucial for the employer to provide adequate training and awareness to employees about such threats. If someone working in a company that doesn't prioritize cybersecurity clicks on 'Enable Login' carelessly, they could put the entire company at risk. If I were an employer, I wouldn't include an 'Enable Login' option."* This statement indicates a comprehensive understanding of how individual actions can affect an organization.

4) Emotional Responses (F4)

Description: Increasing the level of friction leads to intense emotional responses in the users. The emotions range from responsibility, uncertainty, and fear, to feeling patronized.

When presented with increased friction in the phishing warning, users displayed more emotional responses. This theme connects closely to the previous theme, with most emotions arising from being more aware of a conscious choice.

When encountering the increased friction, participants expressed an increased sense of responsibility. Participant 14 explained, *"I can become liable. So ... I'm taking my own risk, too. I'm exposing myself to the threat consciously. And for that [there] might be different consequences."* Participant 17 extended this sentiment to factor in the actual liability in a workplace scenario: *"It's challenging. If you're an employee and fall victim to fraud after clicking 'Enable Login', it's your responsibility, not the company's. This is a complicated situation for an employee... I wouldn't feel comfortable clicking 'Enable Login' because it puts the responsibility solely on me."*

This feeling of responsibility led to uncertainty and fear. She continues: *"I don't have 100% confidence in the IT world ... So, I might be doing something wrong by clicking this ['Enable Login'] button and maybe putting my company and its financial status at risk. There are different aspects of my company at risk - reputational, financial, and many others. I don't feel very comfortable clicking on this button if I'm not 100% sure I can do that."* Introducing the microboundary triggered a thorough evaluation of the risks and her potential liability.

One participant, who had a very mindful stance on security and privacy throughout the interview, felt that the microboundary came across as patronizing. This sentiment overlaps with theme M1 when interpreted as being too obtrusive. He expressed his dismay, saying, *"This would be in the whole patronizing thing, kind of ... removing even further agency from myself ... [The warning is] even more patronizing as compared to the thing before"* (Participant 15).

D. MESSAGE CONTENT

1) Obtrusiveness (M1)

Description: The level of detail of the message influences how obtrusive users perceive the warning to be. Lengthy warnings may be seen as annoying disruptions. Users who consider themselves to be internet-savvy may find detailed explanations patronizing.

Participants generally found short and concise reminders to be more user-friendly. Some participants expressed annoyance with long warning messages, especially if it was shown frequently. They were familiar with the threat and felt the warning was too extensive considering they already knew how to detect phishing without assistance. Participant 14 reflected: *"It depends how many situations I would [see this warning in before] this whole thing gets annoying. It's too much, you know."*

One participant felt more than annoyed at the longer warning, perceiving the detailed explanation in the warning message to be patronizing. He considered himself to be knowledgeable about online risks, and the warnings were designed for less experienced users. He described his sentiment with the analogy of language proficiency levels, feeling that the warnings were designed for a lower proficiency level than his own: *"I feel like I'm C2 in this, and this is for B2 people."* When presented with the shorter, less detailed warning, his negative emotions subsided: *"It's better since it's less patronizing. So it's just kind of nudging me and like reminding me and poking me... It's cleaner, it's simpler."*

2) Effectiveness and Clarity (M2)

Description: The level of detail in the warning message has different implications for its clarity and subsequent effectiveness, depending on the user. Varying levels of information are necessary to make an informed decision in a phishing scenario, at the same time users will gloss over long warning texts.

The data revealed varying preferences among users regarding the level of detail in warning messages. Some responses show that the user's familiarity with phishing and technology impacts their preference for warning message details. Participants with a good grasp of the concept of phishing appreciated the concise, short warning. For them, the cue *"phishing"* is clear and impactful, and further details do not add value.

One participant stated, *"When you say 'Suspected Phishing', people will definitely close it ... it's better than the others"* (Participant 2). Participant 6 phrased their preference very clearly as well, saying, *"because I am familiar with phishing, I don't need all of this other information ... Just show me the URL and warn me."*

In contrast, other participants recognized the short warnings as alerts, but felt they did not provide enough information to understand the exact threat and potential consequences. Participant 17 describes this vague sense of caution with an analogy: *"The computer is saying to me there's a risk. Be careful. Double-check your website ... It's kind of like when you are on the motorway, and you see [an unknown street sign] and you're saying, 'I don't know what it is, but I will be careful.' I will maybe slow down and look around me."*

An interesting admission by some of the participants was that they would skip over lengthy warning messages. This was attributed to a lack of patience and interest in reading through the entire message. As one participant summarized: *"I don't have time to read things"* (Participant 15).

3) Adaptability (M3)

Description: Users attribute different levels of usefulness to an adaptable warning message based on their technical knowledge and experience. Impressions range from 'unnecessary' for technically skilled users to 'very useful' for users who consider their peers.

The adaptability of the warning message was received differently among participants. While some appreciated the option to customize the experience for themselves or different users, others felt it introduced unnecessary complexity.

Participant 18 voiced doubts regarding the perceived complexity introduced by having two modes available: *"Is it worth it? In introducing that complication, I'm not so sure... Is that worth the additional complication to have presented the user with expert mode and non-expert mode?"* He elaborated that if the only difference between the modes was the message content, it might be overcomplicating the extension.

On the other hand, Participant 17 felt that adaptability would increase usability for a diverse audience. She explained that while she herself would prefer the short reminder message, she would value if the more elaborate message was available for colleagues *"if [they are] 60 years old and don't know how to work with a computer, and [they are] on such a website"* This underlines that some users are aware that their fellow workers have different levels of technical understandings and that they might benefit from warnings that are adaptable to their level. This participant pointed out throughout the interview how individuals could affect each other and the entire company, viewing the security mechanism from her own as well as her colleagues' perspectives.

Several participants indicated that they would begin with the elaborate message mode until they were familiarized with the extension, and then potentially switch to the short reminder. Participant 18 described: *"Yeah, I would for sure start with the normal mode and then maybe if I'm getting used to the app... maybe you ... switch over."* Participant 14 elaborated on this, making his decision to switch modes based on how often he would encounter the warning: *"I think it's the frequency of this alert being displayed that makes a difference. And I would opt then to just go to an expert mode ... if this is really like a daily thing."*

One participant related his assessment of the adaptability to the example label it had been given in the mock-up, namely *"Expert Mode"*. He argued that experts do not require a warning at all: *"But then if I'm an expert, then I already know this because I looked at the link... I'm very sorry, but this is redundant because if you are an expert, then you do not need this..."* (Participant 15). This participant would prefer an option to turn the warning mechanism off completely.

V. DISCUSSION

A. TIMING

The timing of warnings was a relevant factor in how users reacted to them. Presenting warnings too early in the phishing attack pipeline may not be effective due to a lack of a sense of threat. When displaying a warning at the last possible moment, users had a better understanding of the risks associated with disregarding the warning. Furthermore, decisions on the timing of warnings can arouse privacy and security concerns among users.

The importance of timing in regard to online security warnings was investigated by Qu et al. [20]. In their scenario, later-stage warnings were more effective at nudging users toward more secure decisions. The results from this research give insights into the effect of timing in the context of active threat phishing warnings. The themes T1 and T2 can be aligned with the findings by Qu et al. [20], and give possible explanations for why later-stage warnings might be more effective at making users more compliant. The themes deal with different levels of perceived security, but they also approach it from different angles. T1 addresses knowledge-based lack of awareness of safety risks at an early stage, whereas T2 describes emotion-centric responses of feeling more safe with early-stage warnings, featuring different levels of concrete understanding of the risks. Users wanted to be shielded from any potential harm as soon as possible, and their discomfort increased with a later warning.

Theme T1 (''Risk Understanding'') illustrates how participants with limited knowledge of phishing were unable to assess the risk associated with following a phishing link, but were more confident about the risks associated with entering your credentials on a phishing page. This suggests that users can have an inadequate mental model of the threat at an earlier stage, which underestimates the potential risks involved. The mental model of the threat is more concise at a later stage. It is crucial to know that there may already be harmful consequences when following the link. Visiting the page may already allow for a malware infection. The phishing page may prompt a download to install malware, or, in combination with a browser vulnerability, trigger a drive-by download, which is an automatic download of malicious software without the

user's knowledge or consent. Furthermore, scripts on the page might attempt to steal browser cookies and saved passwords, which could lead to access to the victim's online accounts. Even in scenarios where no exploit is taking place on the website, clicking the link can inform the attackers that your email address is actively used and potentially receptive to these attacks, which could lead to being targeted in future attacks. Regardless of concrete harm done, ignoring the warning may increase the chances of ultimately becoming a victim of the phishing attack if this early-stage warning is the only warning in the phishing pipeline, and there is no more protection at the later stages. Exposing the users to warnings at all stages of the phishing attack may induce warning fatigue (see Bravo-Lillo et al. [41]), and the user may no longer heed the later stage warnings after having ignored the early stage one.

However, other participants expressed that they would feel safer with an early-stage warning (Theme T2 ''Perceived Safety''). They preferred to avoid coming in contact with a malicious website in the first place. For knowledgeable participants, this sentiment came from being aware of the concrete dangers. For participants who were not concretely aware of the potential risks of following a link, a notable factor in that feeling was uncertainty. Being unsure about the nature of the threat and the consequences made these participants more cautious, despite being uncertain about concrete risks.

Findings T1 and T2 highlight a warning design challenge: Presenting warnings too early in the phishing attack pipeline might not sufficiently highlight the risks, whereas waiting until the user is about to perform a risky action can trigger a more robust risk awareness response. However, presenting the warning right before the credential theft already puts users at risk and can additionally increase discomfort and fear.

A strategy to reconcile between these two findings could be to modify the early-stage warning to be tailored toward those who have a low perception of risk at that stage. The warning at the link attempted to educate the user about the hazard, the potential consequences, and how to avoid the hazard. Theme T2 suggests that a vague sense of danger may already suffice to deter users from proceeding on a potentially dangerous website. An early-stage warning could be combined with a concise message about an exemplary potential risk to invoke a feeling of uncertainty and therefore encourage those users to heed the warning.

Another approach to reconcile these two opposing arguments would be to accommodate the different types of participants. An adaptable warning could present the warning at a suitable timing for the individual user. To implement such an adaptable warning system, it would be essential to identify the factors that would determine the timing of the warning. Possible factors could be the user's technical knowledge or past interactions with warnings.

A striking observation when presenting the participants with a warning in the email inbox was how security-related concerns with the extension were raised. Participants were expressing suspicions about the extension being able to read their emails. Some added additional security concerns were about the extension itself being potentially malicious or introducing a vulnerability. An extension like this would likely always be able to read anything on the user's screen. The baseline warning showed a warning on the phishing website. In order to be able to detect phishing pages, the extension would require permission to scan the content of all websites that are visited. This would include an email client in the browser. However, these concerns were only brought up in the mailbox: The presence of private messages seemed to draw attention to potential privacy issues that could be related to the extension. Emails are regarded as more private than the login website. Concerns about privacy-invading browser extensions are not unfounded. As shown by Carlsson et al. [61], extensions that are given certain permissions can misuse them for malicious purposes.

This observation underscores the importance of designing warning extensions with user privacy in mind. As demonstrated by the participants' reactions, placing warnings in a sensitive context can provoke distrust in the tool. This is particularly concerning for security-enhancing software, which should inspire trust. This implies that designers need to be mindful about the context in which they place warnings, potentially opting for a different timing to display the warning. Overall, it is important to be transparent about the permission the software has, the data it accesses, and how it is used. This could help ease privacy concerns.

B. PLACEMENT

The placement of warnings played a role in the users' experiences with them. The results showed that warnings placed in immediate proximity to the user's current focus were described as more noticeable, but also as more intrusive. Furthermore, the placement was found to be used as an indicator of the source of the message.

The investigation of warning placement was based on the findings of Petelka et al. [18]. They found that placing a warning closer to the phishing link as opposed to placing it in a banner at the top of the email client decreased the click-through rate on phishing links. In our study, the concept of placing the warning close to the threat is applied to a later stage of the attack, namely at the credential theft attempt. Theme P1 (''Visibility'') supports Petelka et al.'s findings, with participants describing the warning next to the login field as more noticeable, leading to greater attention being given to the warning. However, Petelka et al. explain their finding with: ''Our interpretation is that with a banner warning, users search for the suspicious link by hovering over multiple links, hover for a longer time over a phish link, yet might still click the phishing link'' [18, p.12]. In our study, a different perspective for the increased attention surfaced: Participants experience the generalization effect when presented with the warning in the top right corner of the browser. They associate the top right corner of the browser with notifications of low relevance to them, such as browser updates. Compared to regular habituation, which occurs when a person is frequently exposed to the same stimulus, generalization describes the

effect of habituation to one stimulus carrying over to another one [62]. This explains how a habituation-like effect was observed despite the participants seeing the warnings for the first time during the study. Vance et al. [63] investigated how non-essential notifications blurred with security warnings due to generalization. They found that generalization led to decreased attention as well as decreased compliance with warnings. In their paper, they urge developers and designers to design visually unusual security warnings that are resistant to generalization.

The findings from our study suggest that the placement of the warning close to the login field is a way to create such a generalization-resistant warning. However, Themes P2 (''Obtrusiveness'') and P3 (''Uncertainty about the Source of the Warning'') illustrate downsides to this placement.

Theme P2 sheds light on how the login placement of the warning can feel obtrusive and cluttering to the users. This obtrusiveness may even be more pronounced on different login pages than the example presented in the study, which had an empty space next to the login field. In the example, the warning did not cover any information, but that could differ on a website that contains additional elements next to the login section. Another placement-related insight emerged from a key incident when the timing mock-up was presented to Participant 15. The warning location drew his attention to the link and thereby distracted him from the sender address, which surprised the online-security-savvy participant. This introduces a new perspective on the guideline by Petelka et al. [18] of putting the phishing warning where the link is. To address this shortcoming of the warning placement, the sender address could be included and highlighted in the message, similar to how URLs are commonly highlighted as important phishing cues in warnings [56]–[58].

The case of Participant 15 may also have been a product of the artificial environment in which the interview took place. This study's method was not designed for behavioral analysis, and he may not have paid attention to the phishing cue because of the interview taking place. In a natural context, his experience may have been different.

Themes P1 and P2 create tension between visibility and obtrusiveness. While a prominently placed warning may be more effective, it may also be perceived as more annoying. Warning designs need to strike a balance between being visible enough to be effective but not so obtrusive to negatively impact user experience.

Theme P3 raises intriguing points regarding the trust in the warning mechanism. Placing the warning within the website created confusion about the origin of the warning. While some participants unknowingly attributed the warning to the incorrect source (namely the website itself), other participants consciously drew attention to the fact that the source of the warning was not verifiable. A possible explanation for these concerns being raised during the study was that they were primed by the study setup to expect fraud, and were therefore exhibiting increased suspicion. The way the concerns were raised suggests this heightened vigilance, for example, Participant 15 introduced the thought with *"The more I look at this, the more we talk about this, the more I think about this... if the notification is here, then I might think that this is a part of the pop-up."* This level of caution may not reflect typical user interaction.

C. LEVEL OF FRICTION

Exploring the level of friction within phishing warnings provided interesting insights into user perceptions and behaviors. While an increased level of friction successfully heightened user awareness, users with higher proficiency occasionally viewed increased friction as overbearing or unnecessary. Friction also triggered a wide range of emotional responses, from fostering a sense of responsibility to eliciting feelings of patronization.

Two levels of friction were presented: a simple warning reminding users of security risks, and one with a microboundary that interrupted the user's workflow. The concept of security-enhancing friction is inherently based on temporarily creating a negative user experience, as explained by Distler et al. [19]. This has been shown to have a positive effect by disrupting mindless, automatic interaction in crucial situations [39]. The findings in Theme F1 (''Obtrusiveness'') offer insights into how users perceive and accept this mechanism.

A notable aspect of this finding is that obtrusiveness is seen as more acceptable when the data under potential threat is perceived to be significant. For Participant 4, the acceptance of friction was dependent on the perceived sensitivity of the credentials. He did not consider the social media website sensitive enough to warrant the presence of the increased friction. In terms of risk assessment, this participant focused on the immediate risks associated with access to his account's information, but did not mention broader security concerns. These might include extended access to additional accounts in case of password reuse, or the phishing being part of a multi-stage attack, which could endanger not only him but also other users. This has an implication for practical security strategies: In order to increase user acceptance of friction-based warnings, the level of friction may not need to be lowered if the users' risk awareness is enhanced instead. One potential approach would be to incorporate educational information about extended risks into the warning messages. However, the findings from this study suggest that users are not taking warning software seriously in the crucial scenario in which they lack risk awareness. Alternative strategies would include employee training programs or awareness campaigns to educate users about all potential risks.

A key challenge of using security-enhancing friction is user acceptance. If a user finds the mechanism too obtrusive, they may look for ways to circumvent it by uninstalling the software or switching the browser. The friction needs to be adjusted to a level that effectively protects the user but does not alienate them.

Theme F2 (''Diligence and Behavioral Changes'') underlines how friction can influence user behavior. It shows that the microboundary can lead to more diligence and careful

behavior. This aligns with the findings by Cox et al. [39] that a microboundary can disrupt ''autopilot'' behavior and push users into the more deliberate System 2 mode of thinking.

Theme F3 (''Awareness of User Agency'') highlights the increased awareness of user agency when additional friction is introduced. The friction appeared to draw attention to the fact there was a decision to be made, namely to proceed with the login or to go back. The participants' responses reveal a tension between appreciating this agency and feeling overwhelmed by the responsibility it entails. The decision on whether to proceed was the same in all examples presented during the interview. However, with the increased friction, they were reluctant to perform the additional action. It prompted participants to point out that there should not be a user choice at all and that the website should be blocked completely, however, the premise of the warning scenario is that the software cannot accurately discern whether the website is fraudulent or not. This increased awareness of user choice could be leveraged to encourage secure decisions. Theme F2 shows that the microboundary can disrupt the ''autopilot'' behavior, so the user may now be in a System 2 state. Theme F3 provides the additional insight that at that stage, the participants are acutely aware of their agency to make a decision on how to proceed. Depending on how the individual reacts to their agency, the warning could offer more explicit choices to users who value agency, but provide more implicit guidance to those who prefer to avoid such decisions.

Theme F4 (''Emotional Responses'') gives valuable qualitative insights into the emotional reactions of users when confronted with increased friction. The participants' emotional responses ranged from heightened responsibility over fear to feelings of patronization, indicating that the emotional impact of friction can be vastly different for individuals. Heightened emotional responses such as fear can be beneficial for user attention in the context of security warnings. Especially, the feeling of increased responsibility created a basis for safe decision-making. However, increased negative emotions such as fear can create negative user experiences. A particularly intriguing sentiment is the recorded key incident of feeling patronized. Security design revolves around keeping users as safe as possible, which often focuses on novice users as they are the most vulnerable. This is shown by the many Usable Security research publications featuring Johnny, a persona of a novice to average user struggling with cybersecurity [64]–[66]. However, the incident in this study provides a perspective from the other side, exposing a nuanced sentiment of more advanced users. The warnings need to be designed to be effective without making users feel like their autonomy is being undermined. The trade-off between enhancing security and decreasing user experience needs to be appropriate for users of any level of expertise. This predicament makes a strong case for adaptability that adjusts the warning's level of friction depending on the individual user's level of expertise, and possibly other factors (see Section V-E). This would ensure all users are kept safe, while making sure advanced users do not feel patronized.

D. MESSAGE CONTENT

The results regarding the message content highlight how the length and detail of phishing warning messages influence user perception. The reactions were diverse, with some requiring details to understand the threat, and some considering lengthy warnings as disruptions and glossing over them.

Theme M1 (''Obtrusiveness'') mirrors Theme F1 on the topic of Friction (see Section V-C): Increased friction and lengthy warnings had a similar effect in regard to obtrusiveness. M1 also features the key case of feeling patronized by the detailed explanations, paralleling the reaction to the microboundary.

Theme M2 deals with the aspect of effectiveness and clarity of warnings. The effectiveness of warnings is dependent on the message's clarity. Zaaba et al. [6] found that a majority of users dismiss warnings due to the use of technical words and a lack of information provided. In contrast, a study by Lain et al. [16] revealed that more detailed warning messages did not increase the effectiveness of phishing warnings. Our study offers a qualitative perspective on this topic. In the mock-ups, the detailed version of the warning refrained from using technical words and provided further information, whereas the short reminder version used the technical term ''phishing'' and gave no further explanations. Some users prefer short warnings because they are less intrusive (see Theme M1), and because they are perceived as more clear. Users that are already knowledgeable about phishing only require a reminder that this may be a phishing page, and are then able to take appropriate actions without any guidance. But most notably, users are more willing to actually read the message if it is short. This may be the strongest factor supporting and extending Lain et al.'s findings, as the clarity of the content does not matter if the users do not read it. On the other side, some users with less familiarity with phishing did not have a clear understanding of the threat when presented with the shorter warning. Their explanations were in line with Zaaba et al.'s findings regarding the ineffectiveness of warnings with too little information. These findings indicate that both approaches are valid, depending on the concrete audience. While the short reminder may work for some, it may be insufficient for less informed users.

Given the diverse reactions to warning message content, and to the other factors discussed, adaptability emerges as a possible solution for these disparities. Due to the factor-overarching relevance, a subsequent section will be dedicated to adaptability and the discussion of Theme M3 (''Adaptability'').

E. ADAPTABILITY

The concept of adaptability was presented to the participants with a settings section for the browser extension that allowed them to toggle between two modes. Depending on the selected mode, the warning messages would appear with different content.

Lennartsson et al. [17] found in a literature review that adaptability was considered a moderately important part of

Usable Security within the research community. It emerged as the seventh most frequently discussed factor in Usable Security research. The findings in Theme M3 reveal that users perceive adaptability differently, depending on their technical expertise and their consideration of others. While some technically skilled users may deem adaptability unnecessary, others view it as a valuable feature when considering the varied needs of other users.

The insights from M3 suggest that adaptability helps cater to varied user needs, although user acceptance of the feature can vary. Some participants appreciated the option to modify their experience as they grew more accustomed to the tool. While others did not require customization for themselves, they saw value in customization for their colleagues. This underlines that some users are aware that their fellow workers have different levels of technical understanding and that they might benefit from warnings tailored to that level. In a key incident with Participant 17, she described how the security decisions made by individual employees could affect others and even the entire company, displaying a broad understanding of how security goes beyond her own actions. To increase acceptance, it may help to educate users about how adaptable security can increase overall security within a company.

In our findings, a pronounced negative reaction to the detailed message content was recorded. In a key case, Participant 15 felt patronized by detailed warnings (Theme M1), and he felt that no warning was necessary at all for expert users (Theme M3). These findings call attention to the need to consider how to address users with self-perceived high security proficiency. Firstly, self-perceived expertise may not always align with actual security practices [67]. Secondly, despite high expertise, external factors can still make these users vulnerable to threats. Research shows that phishing susceptibility increases with factors like high email load [68], poor sleep quality [69], or being targeted with carefully designed, personalized phishing attacks [70]. A potential solution for this conundrum would be to offer low-intrusion security mechanisms, which are perceived as acceptable to this user group. These security mechanisms could be designed with granular adaptability, allowing for an ''expert mode'' that displays less intrusive reminders than what was presented in the study. This could be text-free, subtle visual cues such as a red backdrop to highlight suspicious URLs, which may be perceived as non-patronizing but still contribute to safe decision-making.

Some participants saw value in the customization option for their own usage. They described their preferred behavior as starting with a detailed message and later switching to a shorter message once they were more familiar with the software or the attacks. This description frames their intended usage as a progression from an ''education mode'' to a ''reminder mode''. This perspective on warning adaptability could be worth exploring. Not only could the framing of a mode as ''educational'' over ''non-expert'' be beneficial to user acceptance and engagement, but educational design best practices could be applied to the warning to deliver a training effect through the warning messages. The tool would then switch to a reminder mode after the user has been educated. These modes would not have to be limited to different message content.

In this study, the factor ''adaptability'' was researched in the context of message content as an example. However, the reactions may have been different if the adaptability had been applied to other factors. Depending on the user and the scenario, different combinations of the configurations could be suitable. This could include configurations outside the range of what was shown to the participants in this study, such as a higher level of friction that requires the user to type in the word ''login'' to proceed. This could also include factors not investigated in this study, as well as some more nuanced differentiation, such as offering more explicit choices to users who value agency, while providing more directive guidance to those who prefer to avoid such decisions (F3).

On the topic of adaptable warnings, it is important to discuss according to which factors the mechanism will adapt. Petelka et al. [18] suggest a context-related approach in which the warning adapts to the nature of the attack, for example depending on whether a phishing message contains a suspicious link or an attachment. In a different approach, the design would be dependent on the user instead of, or in addition to, the threat type. The results of our research have shown that not all users experience the factors of timing, placement, level of friction, and message content in the same way, and an optimal setting has not emerged for any of the factors. In certain cases, like with the discussed usage of an education mode, users may adapt the software themselves. With the configuration left in the hands of the users, however, most users leave settings in their default state [71]. Within an organization, it might make sense for the responsible administrator to decide in which configurations security mechanisms, such as phishing warnings, are displayed. An alternative approach to self- or admin-guided configuration would be automatic adaptiveness. The system could adapt implicitly to individual user characteristics. Derived from our results, these could include technical skill level, online behavior, prior interactions with warnings, past security incidents, and the sensitivity of the threatened data. In a workplace context, this could offer protection to a diverse group of employees.

F. MISUSE POTENTIAL OF FINDINGS

Misuse potential in the findings of this study was only identified on the topic of placement. Despite the increased suspicion observed in participants, potentially influenced by the study's setup, the underlying concern regarding the origin of the warning (P3) is not unfounded. As discussed in Section V-B, the placement of warnings on the website itself rather than as part of the browser's UI, raises legitimate concerns about trust and authenticity. P3 reveals a security issue with the placement. The extension could theoretically be mimicked by adding the element to the website itself. A possible misuse strategy could be to display a green, verifying version of the warning in an attempt to appear more trustworthy. No-

tably for this study, several participants raised these concerns, showing how the placement of the warning not only generates misuse potential, but also mistrust in the software. This issue presents a strong case for a placement that overlaps with the browser UI to allow security-aware users to verify authenticity. The users' trust in security mechanisms is essential for them to heed the warnings. Therefore, the issue of trust should be strongly considered when making design decisions about the placement of warnings.

There is a potential conflict between this conclusion drawn from P3 and the finding that placing warnings in more prominent locations enhances visibility (P1). With legitimate security warnings in more trustworthy but less visible locations, attackers could draw users' attention away from those warnings or real phishing cues and toward fraudulent messages in a high visibility location. This highlights the "dual-use" nature of these findings: while they can be used to enhance the effectiveness of legitimate warnings, they may also create opportunities for attackers to improve deceptive warning. Therefore, a warning design strategy must carefully consider visibility, misuse potential, and user trust.

G. STUDY DESIGN

Participants provided informed consent and were aware that the study examined phishing warnings; the design did not involve deception. The study design replicates an office setting. While being practical and controlled to evaluate the variables of timing, placement, level of friction, and message content (see Table 1 for a comparison with related work), participants' responses may not accurately reflect behaviors and decisions made in a real-world work environment where contextual pressures, task complexity, social dynamics, and the office setting could alter perceptions and interactions. For example, P3 could be non-existent with controlled client software. A long-term study could help to understand the differences. Participants were recruited via flyers on a university campus in Luxembourg, introducing a potential self-selection bias. Moreover, the sample was disproportionately composed of individuals with technical backgrounds. Technical participants might demonstrate different expectations, familiarity, and tolerance for new systems compared to users from non-technical domains. However, if technical participants do not tolerate a new system like phishing warnings, then the system will not succeed. Lastly, the study's geographic limitation to Luxembourg may introduce a cultural bias. Responses collected in this setting may not be representative of behaviors or attitudes in other countries. Following this, a new study could be conducted in geographical different places. Nonetheless, to our knowledge, it is the first study to analyze the four variables of timing, placement, level of friction, and message content. With the qualitative approach, insights were gained that can be used in future studies.

VI. CONCLUSION

The rise in phishing attacks, especially variants such as AI-reinforced and spear phishing, highlights an urgent need for effective countermeasures. When detection software cannot determine with full certainty whether a message is phishing or not, the agency over the assessment has to be placed in the hands of the user. It is crucial that the user is presented with information about the potential threat in a way that will lead to safe decision making. Designing effective and usable phishing warnings is a challenge that requires nuanced understanding of how users experience them. In this study, it was investigated how the factors *timing*, *placement*, *level of friction*, and *message content* influence user experience with browser phishing warnings.

The qualitative investigation reveals that there is no one-size-fits-all approach to phishing warning design, as there are advantages and disadvantages to modifying each design factor. Additionally, the effect of different warning designs is not the same for all users.

This study contributes several key findings. *Timing* modifies the context in which the warning appears, which can inadvertently raise security and privacy concerns by appearing in a more sensitive environment, highlighting the importance of transparency and trust in regard to permissions. *Placement* affects attention and trust. Warnings in locations associated with other notifications may be dismissed due to habituation. Prominent placement can distract users from phishing cues, and ill-considered placement can cause confusion about the warning's source which can even lead to distrust in the security mechanism. *Friction* can increase the user's awareness of their agency and lead to safer decisions. However, high friction can feel patronizing to users with high self-perceived security proficiency, suggesting the need for differentiated approaches. *Message content* must find a balance between clarity and brevity. Long messages are often skipped or perceived as patronizing, while short messages may not fully inform the user. The concept of *adaptability* emerged as a promising solution to the differing individual experiences with warnings. Future research could provide insights on the potential of automatic adaptiveness based on user behavior or contextual factors.

These nuanced findings contribute to the field of Usable Security by expanding the understanding of the individual user perspective on phishing warnings. They can guide further research on user-centric warning design and provide actionable insights for the design of security tools. The themes identified in this research could serve as a foundation for generating questions and options in a survey-type study to allow researchers to generalize and quantify findings. A deceptive lab study or field experiment could investigate the real-life effects of warning design factors. Future studies could also include variables such as color and sound cues, conduct the study on a smartphone, and compare the results with machine learning approaches. Users are the last line of defense against these threats, and designing effective, user-friendly, and individualized defense mechanisms is a crucial step to protect against phishing.

REFERENCES

[1] Sophos Ltd., "Phishing-insights 2021," Sophos Ltd., Abingdon, United Kingdom, Whitepaper, 2021, https://www.sophos.com/resources/phishing-insights-2021.

[2] Anti-Phishing Working Group, "Phishing Activity Trends Report, 4th Quarter 2023," Anti-Phishing Working Group, Lexington, MA, USA, Quarterly Report, Nov. 2023, https://docs.apwg.org/reports/apwg_trends_report_q4_2023.pdf.

[3] ——, "Phishing Activity Trends Report, 1st Quarter 2024," Anti-Phishing Working Group, Lexington, MA, USA, Quarterly Report, May 2024, https://docs.apwg.org/reports/apwg_trends_report_q1_2024.pdf.

[4] B. Schneier, "Semantic attacks: The third wave of network attacks," 2000, [Online]. Available: https://www.schneier.com/essays/archives/2000/10/the-third-wave-of-network-attacks.html. Accessed on: Nov 14, 2025.

[5] A. Prasad and S. Chandra, "PhiUSIIL: A diverse security profile empowered phishing URL detection framework based on similarity index and incremental learning," *Comput. Secur.*, vol. 136, p. 103545, 2024, DOI: 10.1016/j.cose.2023.103545.

[6] Z. F. Zaaba, C. Lim Xin Yi, A. Amran, and M. A. Omar, "Harnessing the Challenges and Solutions to Improve Security Warnings: A Review," *Sens.*, vol. 21, no. 21, p. 7313, Nov. 2021, DOI: 10.3390/s21217313.

[7] P. Boyle and L. A. Shepherd, "MailTrout: A machine learning browser extension for detecting phishing emails," in *Proc. British HCI Conference, virtual*, Jul. 2021, pp. 104–115, DOI: 10.14236/ewic/HCI2021.10.

[8] M. Volkamer, K. Renaud, B. Reinheimer, and A. Kunz, "User experiences of TORPEDO: TOoltip-poweRed Phishing Email DetectiOn," *Comput. Secur.*, vol. 71, pp. 100–113, Nov. 2017, DOI: 10.1016/j.cose.2017.02.004.

[9] M. Cooper, Y. Levy, L. Wang, and L. Dringus, "Heads-up! An Alert and Warning System for Phishing Emails," *Organ. Cybersecur. J. Pract. Process People*, vol. 1, no. 1, pp. 47–68, Jan. 2021, DOI: 10.1108/OCJ-03-2021-0006.

[10] M. Moghimi and A. Y. Varjani, "New Rule-Based Phishing Detection Method," *Expert Sys. Appl.*, vol. 53, pp. 231–242, Jul. 2016, DOI: 10.1016/j.eswa.2016.01.028.

[11] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," *Expert Sys. Appl.*, vol. 117, pp. 345–357, Mar. 2019, DOI: 10.1016/j.eswa.2018.09.029.

[12] O. Bhopen Singh and H. Tahbildar, "A Literature Survey on Anti-Phishing Browser Extensions," *Int. J. Comput. Sci. Eng. Surv.*, vol. 6, no. 4, pp. 21–37, Aug. 2015, DOI: 10.5121/ijcses.2015.6402.

[13] A. A. Zuraiq and M. Alkasassbeh, "Review: Phishing Detection Approaches," in *Proc. ICTCS, Como, Italy*, Oct. 2019, pp. 1–6, DOI: 10.1109/ICTCS.2019.8923069.

[14] V. Shahrivari, M. M. Darabi, and M. Izadi, "Phishing Detection Using Machine Learning Techniques," Sep. 2020, [Online]. DOI: 10.48550/arXiv.2009.11116. Accessed on: May 26, 2025.

[15] S. Das, A. Kim, Z. Tingle, and C. Nippert-Eng, "All about phishing: Exploring user research through a systematic literature review," in *Proc. HAISA, Nicosia, Cyprus*, Aug. 2019, pp. 189–202.

[16] D. Lain, K. Kostiainen, and S. Čapkun, "Phishing in Organizations: Findings from a Large-Scale and Long-Term Study," in *Proc. SP, San Francisco, CA, USA*, May 2022, pp. 842–859, DOI: 10.1109/SP46214.2022.9833766.

[17] M. Lennartsson, J. Kävrestad, and M. Nohlberg, "Exploring the Meaning of "Usable Security"," in *Proc. HAISA, Mytilene, Greece*, ser. IFIP Advances in Information and Communication Technology, N. Clarke and S. Furnell, Eds., 2020, pp. 247–258, DOI: 10.1007/978-3-030-57404-8_19.

[18] J. Petelka, Y. Zou, and F. Schaub, "Put Your Warning Where Your Link Is: Improving and Evaluating Email Phishing Warnings," in *Proc. CHI, Glasgow, Scotland, United Kingdom*, May 2019, pp. 1–15, DOI: 10.1145/3290605.3300748.

[19] V. Distler, G. Lenzini, C. Lallemand, and V. Koenig, "The Framework of Security-Enhancing Friction: How UX Can Help Users Behave More Securely," in *Proc. NSPW, online*, Oct. 2020, pp. 45–58, DOI: 10.1145/3442167.3442173.

[20] L. Qu, R. Xiao, and W. Shi, "Interactions of Framing and Timing in Nudging Online Game Security," *Comput. Secur.*, vol. 124, p. 102962, Jan. 2023, DOI: 10.1016/j.cose.2022.102962.

[21] B. Naqvi, K. Perova, A. Farooq, I. Makhdoom, S. Oyedeji, and J. Porras, "Mitigation strategies against the phishing attacks: A systematic literature review," *Comput. Secur.*, vol. 132, p. 103387, Sep. 2023, DOI: 10.1016/j.cose.2023.103387.

[22] H. Siadati, S. Palka, A. Siegel, and D. McCoy, "Measuring the effectiveness of embedded phishing exercises," in *Proc. CSET, Vancouver, BC, Canada*, 2017, pp. 1–8.

[23] P. Kumaraguru, S. Sheng, A. Acquisti, L. F. Cranor, and J. Hong, "Lessons from a Real World Evaluation of Anti-Phishing Training," in *Proc. eCrime, Atlanta, GA, USA*, Oct. 2008, pp. 1–12, DOI: 10.1109/ECRIME.2008.4696970.

[24] D. Jampen, G. Gür, T. Sutter, and B. Tellenbach, "Don't Click: Towards an Effective Anti-Phishing Training. A Comparative Literature Review," *Hum.-centric Comput. Inf. Sci.*, vol. 10, no. 1, p. 33, Aug. 2020, DOI: 10.1186/s13673-020-00237-7.

[25] G. Misra, N. A. G. Arachchilage, and S. Berkovsky, "Phish Phinder: A Game Design Approach to Enhance User Confidence in Mitigating Phishing Attacks," in *Proc. HAISA, Adelaide, Australia*, 2017, pp. 41–51.

[26] F. Rubia, A. Yasin, and L. Liu, "How Persuasive Is a Phishing Email? A Phishing Game for Phishing Awareness," *J. Comput. Secur.*, vol. 27, no. 6, pp. 581–612, Jan. 2019, DOI: 10.3233/JCS-181253.

[27] Z. Wen, Z. Lin, R. Chen, and E. Andersen, "What.Hack: Engaging Anti-Phishing Training Through a Role-playing Phishing Simulation Game," in *Proc. CHI, Glasgow, Scotland, UK*, Apr. 2019, p. 12, DOI: 10.1145/3290605.3300338.

[28] X. Chen, M. Sacré, G. Lenzini, S. Greiff, V. Distler, and A. Sergeeva, "The Effects of Group Discussion and Role-playing Training on Self-efficacy, Support-seeking, and Reporting Phishing Emails: Evidence from a Mixed-design Experiment," in *Proc. CHI, Honolulu, HI, USA*, 2024, DOI: 10.1145/3613904.3641943.

[29] S. Bell and P. Komisarczuk, "An analysis of phishing blacklists: Google Safe Browsing, OpenPhish, and PhishTank," in *Proc. ACSW, Melbourne, VIC, Australia*, 2020, pp. 1–11, DOI: 10.1145/3373017.3373020.

[30] R. Srinivasa Rao and A. R. Pais, "Detecting Phishing Websites using Automation of Human Behavior," in *Proc. CPSS, Abu Dhabi, United Arab Emirates*, 2017, p. 33–42, DOI: 10.1145/3055186.3055188.

[31] P. Bountakas, K. Koutroumpouchos, and C. Xenakis, "A Comparison of Natural Language Processing and Machine Learning Methods for Phishing Email Detection," in *Proc. ARES, Vienna, Austria*, 2021, pp. 1–12, DOI:10.1145/3465481.3469205.

[32] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems," *Heliyon*, vol. 5, no. 6, p. e01802, Jun. 2019, DOI: 10.1016/j.heliyon.2019.e01802.

[33] D. Akhawe and A. P. Felt, "Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness," in *Proc. USENIX Security, Washington, D.C., USA*, 2013, pp. 257–272.

[34] R. W. Reeder, A. P. Felt, S. Consolvo, N. Malkin, C. Thompson, and S. Egelman, "An experience sampling study of user reactions to browser warnings in the field," in *Proc. CHI, Montreal, QC, Canada*, Apr. 2018, pp. 1–13, DOI: 10.1145/3173574.3174086.

[35] A. Aggarwal, A. Rajadesingan, and P. Kumaraguru, "PhishAri: Automatic realtime phishing detection on Twitter," in *Proc. eCrime, Las Croabas, PR, USA*, 2012, pp. 1–12, DOI: 10.1109/eCrime.2012.6489521.

[36] F. Greco, G. Desolda, P. Buono, and A. Piccinno, "Enhancing Phishing Defenses: The Impact of Timing and Explanations in Warnings for Email Clients," *Comput. Standards Interfaces*, vol. 93, p. 103982, 2025, DOI: 10.1016/j.csi.2025.103982.

[37] S. Bender, S. Horn, G. Loewenstein, and O. Roberts, "Phishing feedback: just-in-time intervention improves online security," *Behav. Public Policy*, p. 1–13, 2024, DOI: 10.1017/bpp.2024.19.

[38] J. Petelka, B. Berens, C. Sugatan, M. Volkamer, and F. Schaub, "Restricting the Link: Effects of Focused Attention and Time Delay on Phishing Warning Effectiveness ," in *Proc. IEEE SP, San Francisco, CA, USA*. Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 1–19. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/SP61157.2025.00007

[39] A. L. Cox, S. J. Gould, M. E. Cecchinato, I. Iacovides, and I. Renfree, "Design Frictions for Mindful Interactions: The Case for Microboundaries," in *Proc. CHI EA, San Jose, CA, USA*, May 2016, pp. 1389–1397, DOI: 10.1145/2851581.2892410.

[40] E. J. Slifkin and M. B. Neider, "Phishing interrupted: The impact of task interruptions on phishing email classification," *Int. J. Hum.-Comp. Stud.*, vol. 174, p. 103017, Jun. 2023, DOI: 10.1016/j.ijhcs.2023.103017.

[41] C. Bravo-Lillo, L. Cranor, S. Komanduri, S. Schechter, and M. Sleeper, "Harder to Ignore? Revisiting Pop-Up Fatigue and Approaches to Prevent It," in *Proc. SOUPS, Menlo Park, CA, USA*, 2014, pp. 105–111.

[42] C. Bravo-Lillo, S. Komanduri, L. F. Cranor, R. W. Reeder, M. Sleeper, J. Downs, and S. Schechter, "Your attention please: Designing security-decision UIs to make genuine risks harder to ignore," in *Proc. SOUPS, Newcastle, United Kingdom*, Jul. 2013, pp. 1–12, DOI: 10.1145/2501604.2501610.

[43] A. Dennis and R. Minas, "Security on Autopilot: Why Current Security Theories Hijack Our Thinking and Lead Us Astray," *ACM SIGMIS Database: DATABASE Adv. Inf. Syst.*, vol. 49, pp. 15–38, Apr. 2018, DOI: 10.1145/3210530.3210533.

[44] D. Kahneman, *Thinking, Fast and Slow*, ser. Penguin Psychology. London, United Kingdom: Penguin Books, 2012.

[45] M. Harbach, S. Fahl, P. Yakovleva, and M. Smith, "Sorry, I don't get it: An analysis of warning message texts," in *Financial Cryptography and Data Security*, D. Hutchison, T. Kanade, and A. A. Adams, Eds. Berlin, Heidelberg, Germany: Springer Berlin Heidelberg, 2013, vol. 7862, pp. 94–111, DOI: 10.1007/978-3-642-41320-9_7.

[46] Z. F. Zaaba and T. K. Boon, "Examination on Usability Issues of Security Warning Dialogs," *J. Multidiscip. Eng. Sci. Technol. (JMEST)*, vol. 2, no. 6, pp. 26–35, Jun. 2015.

[47] O. Sarker, A. Jayatilaka, S. Haggag, C. Liu, and M. A. Babar, "A multivocal literature review on challenges and critical success factors of phishing education, training and awareness," *J. Syst. Softw.*, vol. 208, p. 111899, Feb. 2024, DOI: 10.1016/j.jss.2023.111899.

[48] V. Braun and V. Clarke, "Using Thematic Analysis in Psychology," *Qual. Res. Psychol.*, vol. 3, no. 2, pp. 77–101, Jan. 2006, DOI: 10.1191/1478088706qp063oa.

[49] Ustuk, Özgehan, "Thematic analysis with maxqda: Step-by-step guide," 2025, [Online]. Availabe: https://www.maxqda.com/blogpost/thematic-analysis-with-maxqda-step-by-step-guide. Accessed on: Nov 14, 2025.

[50] K. Krol, M. Moroz, and M. A. Sasse, "Don't Work. Can't Work? Why It's Time to Rethink Security Warnings," in *Proc. CRiSIS, Cork, Ireland*, Oct. 2012, pp. 1–8, DOI: 10.1109/CRISIS.2012.6378951.

[51] Figma, "Figma: The collaborative interface design tool," 2025, [Online]. Availabe: https://www.figma.com/. Accessed on: Nov 14, 2025.

[52] LinkedIn, "LinkedIn," 2025, [Online]. Availabe: https://www.linkedin.com/. Accessed on: Nov 14, 2025.

[53] Check Point Research Team, "LinkedIn still number one brand to be faked in phishing attempts while Microsoft surges up the rankings to number two spot in Q2 report," Jul. 2022, [Online]. Availabe: https://blog.checkpoint.com/2022/07/19/linkedin-still-number-one-brand-to-be-faked-in-phishing-attempts-while-microsoft-surges-up-the-rankings-to-number-two-spot-in-q2-report/. Accessed on: Nov 14, 2025.

[54] A. Black, P. Luna, O. Lund, and S. Walker, *Information Design: Research and Practice*. Milton Park, United Kingdom: Taylor & Francis, Jan. 2017.

[55] Mozilla, "Mozilla phishing test site," 2025, [Online]. Availabe: https://www.itisatrap.org/firefox/its-a-trap.html. Accessed on: Nov 14, 2025.

[56] Netcraft, "Netcraft," 2025, [Online]. Availabe: https://www.netcraft.com/. Accessed on: Nov 14, 2025.

[57] ZecOps Research Team, "Introducing ZecOps Anti-Phishing Extension," Apr. 2021, [Online]. Available: https://blog.zecops.com/announcements/introducing-zecops-anti-phishing-extension/. Accessed on: Nov 14, 2025.

[58] M. Volkamer, "TORPEDO," Feb. 2022, [Online]. Available: https://secuso.aifb.kit.edu/TORPEDO.php/. Accessed on: Nov 14, 2025.

[59] B. Ramkumar, "PicoPalette phishing site detector plugin," 2018, [Online]. Availabe: https://github.com/picopalette/phishing-detection-plugin. Accessed on: Nov 14, 2025.

[60] M. Hennink and B. N. Kaiser, "Sample sizes for saturation in qualitative research: A systematic review of empirical tests," *Soc. Sci. Med.*, vol. 292, p. 114523, 2022, DOI: 10.1016/j.socscimed.2021.114523.

[61] R. Carlsson, S. Rauti, and T. Heino, "A Case Study of a Privacy-Invading Browser Extension," in *Proc. ICITS, Cusco, Peru*, ser. Lecture Notes in Networks and Systems, Á. Rocha, C. Ferrás, and W. Ibarra, Eds., 2023, pp. 127–134, DOI: 10.1007/978-3-031-33258-6_12.

[62] C. H. Rankin, T. Abrams, R. J. Barry, S. Bhatnagar, D. F. Clayton, J. Colombo, G. Coppola, M. A. Geyer, D. L. Glanzman, S. Marsland, F. K. McSweeney, D. A. Wilson, C.-F. Wu, and R. F. Thompson, "Habituation Revisited: An Updated and Revised Description of the Behavioral Characteristics of Habituation," *Neurobiol. Learn. Mem.*, vol. 92, no. 2, pp. 135–138, Sep. 2009, DOI: 10.1016/j.nlm.2008.09.012.

[63] A. Vance, J. L. Jenkins, and B. B. Anderson, "The fog of warnings: How non-essential notifications blur with security warnings," in *Proc. SOUPS, Santa Clara, CA, USA*, 2019, pp. 407–420.

[64] S. Das, A. Dingman, and L. J. Camp, "Why Johnny doesn't use two factor a two-phase usability study of the FIDO U2F security key," in *Financial Cryptography and Data Security*, S. Meiklejohn and K. Sako, Eds. Berlin, Heidelberg, Germany: Springer Berlin Heidelberg, 2018, vol. 10957, pp. 160–179, DOI: 10.1007/978-3-662-58387-6_9.

[65] Z. Benenson, G. Lenzini, D. Oliveira, S. Parkin, and S. Uebelacker, "Maybe Poor Johnny Really Cannot Encrypt: The Case for a Complexity Theory for Usable Security," in *Proc. NSPW, Twente, Netherlands*, Sep. 2015, pp. 85–99, DOI: 10.1145/2841113.2841120.

[66] G. Cybenko, "Why Johnny can't evaluate security risk," *IEEE Secur. Priv.*, vol. 4, no. 01, pp. 5–5, Jan. 2006, DOI: 10.1109/MSP.2006.30.

[67] R. Wash and E. Rader, "Too Much Knowledge? Security Beliefs and Protective Behaviors Among United States Internet Users," *Proceedings fo the 11th Symposium On Usable Privacy and Security (SOUPS)*, pp. 309–325, 2015, https://www.usenix.org/conference/soups2015/proceedings/presentation/wash.

[68] E. Rozentals, "Email Load and Stress Impact on Susceptibility to Phishing and Scam Emails," Master's Thesis, Department of Computer Science, Electrical and Space Engineering, Digital Services and Systems, Luleå University of Technology, Luleå, Sweden, 2021.

[69] H. Qahri-Saremi and O. Turel, "Situational Contingencies in Susceptibility of Social Media to Phishing: A Temptation and Restraint Model," *J. Manag. Inf. Sys.*, vol. 40, no. 2, pp. 503–540, Apr. 2023, DOI: 10.1080/07421222.2023.2196779.

[70] Z. Benenson, F. Gassmann, and R. Landwirth, "Unpacking Spear Phishing Susceptibility," in *Financial Cryptography and Data Security*, M. Brenner, K. Rohloff, J. Bonneau, A. Miller, P. Y. Ryan, V. Teague, A. Bracciali, M. Sala, F. Pintore, and M. Jakobsson, Eds. Cham, Switzerland: Springer International Publishing, 2017, vol. 10323, pp. 610–627, DOI: 10.1007/978-3-319-70278-0_39.

[71] Y. Wang and Y.-W. Mo, "The Effect of Default Options on Consumer Decisions in the Product Configuration Process," in *Proc. ConfWS, Graz, Austria*, ser. CEUR Workshop Proceedings, Jun. 2018, pp. 31–36.

STEFANIE PHAM received the B.Sc. degree in Media Informatics in 2020 and the M.Sc. degree in Computer Science in 2023 from Ludwig Maximilian University of Munich, Germany. Her academic research focused on social engineering, cryptographic security, and usable security, with projects conducted in collaboration with the University of Luxembourg and the University of the Bundeswehr Munich. She is currently working as a software engineer in the financial technology sector. Her professional interests include back-end development and secure software systems.

GABRIELE LENZINI is Associate Professor at the University of Luxembourg, Chief Scientist II at the Interdisciplinary Center for Security Reliability and Trust (SnT), and head of the Interdisciplinary Research group in Sociotechnical Cybersecurity (IRiSC). His expertise is in the analysis and design of security solutions, particularly systems where computer science meets other disciplines, such as social sciences, physics, and law. He holds a PhD in Computer Science (2005, University of Twente, The Netherlands). He participated in the development and execution of numerous national and international projects, several with addressing problem of privacy and security in socio-technical systems.

DANIELA PÖHN is a senior researcher at the Research Institute Cyber Defence and Smart Data (RI CODE) and the University of the Bundeswehr Munich in Neubiberg, Germany. Her research is primarily focused on identity management and social engineering. In her role as a research assistant at the Leibniz Supercomputing Centre, she was active in the EU GÉANT project and doing her doctorate in parallel at Ludwig-Maximilians-Universität München on the subject of dynamic identity management in federations. She did her habilitation at the University of the Bundeswehr Munich.