



A Multi-vocal Literature Review on challenges and critical success factors of phishing education, training and awareness[☆]

Orvila Sarker^{a,b,d,*}, Asangi Jayatilaka^{a,b}, Sherif Haggag^b, Chelsea Liu^c, M. Ali Babar^{a,b,d}

^a Centre for Research on Engineering Software Technologies (CREST), The University of Adelaide, Australia

^b School of Computer and Mathematical Sciences, The University of Adelaide, Australia

^c Adelaide Business School, The University of Adelaide, Australia

^d Cyber Security Cooperative Research Centre, Australia

ARTICLE INFO

Keywords:

Phishing education
Phishing training
Phishing awareness
Human factors
Human-centric security
Usable security

ABSTRACT

Background: Phishing is a malicious attempt by cyber attackers to steal personal information through deception. Phishing attacks are often aided by carefully crafted phishing emails, which can go undetected by automated anti-phishing tools due to their limited accuracy. Studies found that user education, training, and awareness can thwart phishing attacks. Understanding diverse interconnected challenges and critical success factors of phishing education, training, and awareness (PETA) approaches can help improve organizations' defense against phishing.

Objective: This study presents a comprehensive, structured view of the challenges and critical success factors of the design, implementation, and evaluation stages of PETA.

Method: We have conducted a Multi-vocal Literature Review (MLR) by systematically collecting 53 academic studies and 16 grey studies from popular databases by following a well-known MLR guideline.

Results: We identified 20 challenges and 23 critical success factors, some of which involve human-centric and socio-technical factors in PETA. Our findings point out the need for designing explainable anti-phishing systems and developing automated tools and platforms to conduct real-world phishing studies.

Conclusion: Our systematic analysis of 69 studies has enabled us to highlight the need for addressing human-centric issues, incorporating users' knowledge gaps, and adopting personalized approaches in PETA.

1. Introduction

Phishing is a form of cyber attack where perpetrators attempt to steal users' personal information through deception (Jenkins et al., 2022). The second quarter of 2022 is considered the worst quarter for phishing in history with nearly 1,097,811 phishing attacks and with a significant increase of smishing¹ and vishing² (APWG, 2022). According to a global survey conducted in 2021 with 351 practitioners and decision-makers in large organizations across Asia/Pacific, Europe, North America, and the Middle East, organizations are experiencing financial losses, loss of customer trust, customer data breaches, loss of intellectual property, damage of company reputation, loss of investor confidence, regulator penalties and disruptions of business operations due to phishing attacks (Cofense, 2022; Szathmari, 2019). These significant negative consequences of phishing attacks give rise to the urgency for organizations to safeguard themselves against this substantial risk.

One form of protection is offered by anti-phishing software designed to detect phishing attempts and prevent them from reaching end-users. However, given that automated anti-phishing defense cannot guarantee blanket effectiveness, end-users themselves constitute the last line of defense against phishing attempts (Schechter et al., 2007; Arachchilage et al., 2016; Petelka et al., 2019; Chen et al., 2018). Specifically, many automated anti-phishing solutions for detecting phishing have been proposed (e.g., Alani and Tawfik (2022), Magdy et al. (2022), Liu et al. (2022)). These solutions typically rely on machine learning or probabilistic algorithms and suffer from false positives and false negatives (for example, PhishNot proposed by Alani and Tawfik (2022) results in 2.18% false positive results and 3.22% false negative results). Consequently, automated anti-phishing solutions cannot be considered a comprehensive solution for preventing phishing attacks, prompting organizations to move towards more human-centric approaches to minimize their risks and susceptibility to phishing.

[☆] Editor: Burak Turhan.

* Corresponding author at: Centre for Research on Engineering Software Technologies (CREST), The University of Adelaide, Australia.
E-mail address: orvila.sarker@adelaide.edu.au (O. Sarker).

¹ Practice of utilizing SMS to initiate phishing (Desolda et al., 2021).

² Use of phone calls as attack vector to initiate phishing (Desolda et al., 2021).

Table 1

Focus areas of this MLR.

Source: Adopted from Desolda et al. (2021) and Hu et al. (2021).

Topic	Definition
Phishing education	Education interventions are aimed at developing user knowledge and skills of recognizing phishing attempts, enhancing user understanding, and enabling users to gain insights about phishing.
Phishing training	The main goal of training interventions is to provide users with interactive and practical experience about how to protect themselves from phishing attempts.
Phishing awareness	Awareness interventions are commonly deployed as interruptions to users' regular workflow, to provide warnings and raise awareness about possible occurrences of phishing attacks. Awareness interventions also provide users with design choices to detect and identify phishing attacks (e.g., custom icon, trust logo, sender highlighting).

Despite an increasing number of available anti-phishing interventions (e.g., browser phishing warning (Petelka et al., 2019), anti-phishing game (CJ et al., 2018), anti-phishing training (Sjouwerman, 2021)), end-users still fall prey to phishing attacks. This is because the success of anti-phishing interventions in changing end-user behavior can vary substantially depending on their design, implementation, and evaluation (Franz et al., 2021; Kumaraguru et al., 2007; Sheng et al., 2007; Caballero, 2017). Therefore, effective design, implementation, and evaluation of anti-phishing interventions are key to enabling organizations to turn their employees from a potential source of cyber security vulnerability into their strongest line of defense, by providing employees with the skills to identify and report phishing attacks (Jayatilaka et al., 2021; Bada et al., 2019).

In order to offer effective anti-phishing interventions, researchers, and practitioners need to possess an understanding and awareness of the challenges and critical success factors associated with the design, implementation, and evaluation stages. This creates a significant need for a synthesized body of knowledge regarding the determinants of the effectiveness of anti-phishing interventions. Yet despite the rapidly growing academic research into anti-phishing interventions, a vast majority of prior studies focus on specific, individual success factors or challenges, thereby creating a vast and scattered body of evidence, which would be difficult for practitioners to comb through when designing their anti-phishing interventions. To the best of our knowledge, no existing study has conducted a review aimed at organizing this body of knowledge on anti-phishing education, training, and awareness interventions to provide a comprehensive and in-depth synthesis of the existing evidence of factors that drive the effectiveness of anti-phishing interventions. Responding to this lack of investigation into an important topic and the critical need for a one-stop shop where a synthesized body of knowledge can be accessed, we conduct a Multi-Vocal Review (MLR) by systematically analyzing the peer-reviewed and grey literature on this topic.

Conducting an MLR offers several advantages: (i) provides richer data and strong ecological validity in contrast to laboratory settings used in academic studies as indicated by Greene et al. (2018); (ii) allows a more comprehensive analysis for answering the relevant research questions by combining the state-of-the-art and the state-of-practice (Tom et al., 2013); (iii) given that anti-phishing interventions are inherently an industry-oriented practice, including the voice of practitioners ensure that practitioners' experience and industry viewpoints are not missed (Benzies et al., 2006; Mahood et al., 2014); (iv) enables to canvas of abundant practical information from diverse document sources, such as (for example, phishing vendor manuals and guides to run phishing campaigns and evaluate phishing simulations (Terranova Security, 2022)), which provide insights into real-world policies and practices.

Table 2

Comparison of our study with the related existing studies.

Contribution	Jampen et al. (2020)	Franz et al. (2021)	Our study
Challenges in PETA	–	–	✓
Critical factors in design, implementation, and evaluation phases of PETA	–	–	✓
Factors having impact on anti-phishing training	✓	–	✓
Phishing attack vector	–	✓	–
Time for PETA intervention reception	–	✓	–
Taxonomy of the existing solutions on user education, training, and awareness	–	✓	–
Future research directions	✓	–	✓
Study type	Survey	SLR	MLR

– : topic not discussed, ✓ : topic discussed.

This study makes a number of significant contributions:

- We provide an in-depth analysis of the challenges in phishing education, training, and awareness (PETA) interventions to enable researchers and practitioners to gain a better understanding of the limitations of anti-phishing initiatives, which in turn helps improve their effectiveness in safeguarding organizations against future phishing attacks.
- We offer a comprehensive overview of critical factors in the design, implementation, and evaluation stages that determine the success of PETA interventions, providing researchers and practitioners with novel insights and guidance on how to enhance the success of anti-phishing initiatives in complex real-world contexts.
- We present a set of recommendations to guide researchers and practitioners, based on prior empirical evidence, to develop novel approaches to PETA to counter phishing attacks.

2. Related work

Anti-PETA interventions (the terms education, training, and awareness are defined in Table 1) constitute an essential line of defense to mitigate phishing threats that bypass automated detection tools (Sheng et al., 2007; Heartfield and Loukas, 2015). The importance of anti-phishing interventions has attracted significant attention from researchers to conduct human-centric phishing studies. Nevertheless, despite the large and scattered body of evidence on PETA interventions, the prior research has yet to provide a comprehensive overview of the challenges and critical factors that determine the success of PETA interventions. Furthermore, the utmost importance of real-world industrial settings in investigating phishing was emphasized in several studies (e.g., Althobaiti et al. (2021), Burda et al. (2020)). Replicating phishing studies to evaluate the effects of variations in different industrial settings was also recommended (Burda et al., 2020). Despite being an industry-oriented domain, the inclusion of the practitioner perspective has been overlooked in previous reviews. Consequently, a comprehensive understanding of the prevailing challenges and essential factors for successful outcomes remains incomplete. This study aims to address this research gap by synthesizing a comprehensive body of knowledge derived from practical experiences in the industrial setting. This endeavor will encompass insights obtained from diverse sources of grey literature, effectively capturing the perspectives of practitioners.

A body of the existing work (Franz et al. (2021), Jampen et al. (2020)) has attempted to provide an overview of the existing evidence including the taxonomy of user-based PETA interventions, discussed various elements pertaining to the training materials. While the utilization of a taxonomy discussed by Franz et al. (2021) aids in comprehending the existing interventions and their underlying mechanisms, as well

Table 3
Research questions and motivations.

Research Questions	Motivations
RQ1. What are the challenges for designing, implementing, and evaluating phishing training, education, and awareness interventions?	The motivation of this RQ is to understand the constraints and obstacles faced by researchers and practitioners in educating, training, and raising user awareness about phishing during the design, implementation, and evaluation stages.
RQ2. What are the critical success factors for designing, implementing, and evaluating phishing training, education, and awareness interventions?	The primary motivation is to identify the potentially influential factors that can provide actionable insights to the researchers and practitioners to develop improved phishing training, education, and awareness interventions.

as facilitating comparative analysis, it offers limited insight into the challenges faced by these interventions. A comprehensive analysis of the prevailing challenges remains absent in the literature. A thorough comprehension of these challenges can greatly assist researchers and practitioners in devising more effective solutions to address or enhance the current issues within these interventions. The purpose of our study is to bridge this gap by conducting a comprehensive investigation. Our investigation deviates from the study conducted by [Jampen et al. \(2020\)](#), which primarily delved into phishing training interventions, by examining a range of factors that influence the efficacy of phishing training. Our study, in contrast, concentrates on the challenges and pivotal elements that ascertain the effectiveness of PETA encompassing three distinct interventions: phishing education, training, and awareness.

The examination of user demographic information, methodology and evaluation techniques, and various human factors as presented by [Baki and Verma \(2021\)](#), [Hu et al. \(2021\)](#), [Das et al. \(2019\)](#), and [Desolda et al. \(2021\)](#) has made a significant contribution to the existing body of knowledge. These scholarly works have provided valuable insights into the diverse characteristics and vulnerabilities of distinct user groups, thereby benefiting both researchers and practitioners. Nonetheless, for this knowledge to effectively inform the development of targeted and customized intervention strategies, a comprehensive overview of the specific challenges encountered by the end-users and the factors that contribute to successful interventions is necessary. It is the objective of our study to address this gap by presenting an inclusive analysis of the challenges faced and critical success factors associated with these interventions. The findings we report can aid intervention designers in tailoring educational materials, training programs, and communication strategies to effectively engage and educate diverse user groups.

[Table 2](#) summarizes the specific areas of contribution offered by our study in comparison to these prior studies.

3. Research methodology

We follow the guidelines provided by [Kitchenham and Charters \(2007\)](#) and [Garousi et al. \(2019\)](#) to conduct our MLR. Our MLR aims to provide a comprehensive overview of the challenges and critical success factors in the design, implementation, and evaluation stages of PETA. [Table 3](#) shows the research questions and motivations of our MLR.

3.1. Search strategy

3.1.1. Academic literature

To collect academic studies, we used Scopus³ as our search database. The decision to employ Scopus search engine to identify the relevant

primary studies was based on: (i) the experiences reported by several other studies ([Shahin et al., 2017](#); [Kitchenham et al., 2010](#); [Zahedi et al., 2016](#); [Shahin et al., 2020](#)) justifying that Scopus indexes a large majority of the journals and conference papers in indexed by many other search engines such as IEEE Xplore, ACM Digital Library and SpringerLink; (ii) the fact that Scopus track a large number of journals and conferences in software engineering and computer science ([Shahin et al., 2017](#); [Kitchenham et al., 2010](#); [Zahedi et al., 2016](#); [Shahin et al., 2020](#)) that were the main target of this review; and (iii) our pilot search results confirming the comprehensiveness of Scopus compared to other databases by verifying that no important studies were overlooked by solely relying on Scopus.

We conducted pilots with other databases such IEEE Xplore and ACM Digital Library. The search in IEEE Xplore yielded a total of 888 studies, after removing 52 duplicates. From this sample, 70 studies were randomly selected and cross-checked with Scopus search results. Of these, 64 studies were found in Scopus and the remaining 6 were not relevant to the study. Subsequently, 40 studies were identified from the ACM Digital Library, of which 25 were randomly selected and cross-checked with Scopus. Among these, 22 studies were found in Scopus, and the remaining 3 were not relevant to the study. Hence based on the pilot search we can conclude that no relevant studies are missed by using Scopus as the search engine. As a result, we are confident that using Scopus as the search engine could cover most of the relevant papers for this study. A similar observation was made by [Kitchenham et al. \(2010\)](#) in their study. By comparing the results of Scopus with a manual search in their study, [Kitchenham et al. \(2010\)](#) observed that Scopus covered all the relevant papers that used appropriate terminology.

We refined our search queries by conducting two preliminary searches. In the first search, we reviewed highly-ranked studies (CHI, SOUPS, IEEE S&P) and search queries used in previous surveys (discussed in Section 2) related to the scope of our research. To develop our search queries, we used the term “phish*” consistently and combined it with alternative terms from a predefined list that reflected our study’s focus. These alternative terms included “train*”, “awar*”, “educat*”, “teach*”, “learn*”, “interven*”, “nudge*”, and “warn*”.

In the initial phase of our search, we found that the above-mentioned keywords did not generate sufficient user-based studies that offered guidance or assistance to users in combating phishing. To include potential relevant studies, we conducted a second pilot search. This phase of the search was more complex due to inconsistent terminology used in the literature, as explained by [Franz et al. \(2021\)](#). To identify studies closely aligned with the scope of our research, we used keywords such as “security indicators”, “alert*”, and “protect*”, which resulted in more comprehensive and relevant findings. We examined the titles, abstracts, and keywords of previous studies to perform our search.

3.1.2. Grey literature

We choose Google⁴ as a search engine to collect grey literature, like numerous prior studies ([Garousi et al., 2017](#); [Garousi and Mäntylä, 2016](#); [Islam et al., 2019](#); [Butijn et al., 2020](#)). Given the voluminous nature and complexity of grey literature in comparison with academic literature, we streamlined our search keywords when searching grey literature by including only the search terms “education”, “training”, “awareness” alongside the search term “phish*” when collecting grey literature, to produce more targeted and relevant results.

Google’s method of searching is different from scholarly databases like Scopus. Google searches the specified search terms across all indexed pages ([Cascavilla et al., 2021](#)); As a result, we observed that it yielded irrelevant outcomes when the same set of keywords employed for Scopus was used for the Google search. For instance, when using the search terms “protect*” and “learn*”, Google returned outcomes

³ <https://www.scopus.com/>

⁴ <https://www.google.com/>

Table 4
Conceptualization of search strings.
(search strings are categorized based on the definition provided in Table 1).

Study	Focus areas	Search Strings	Source
Academic	Phishing Education Training Awareness	“phish*” AND (“educat*” OR “teach*” OR “learn*”) AND “train*” AND (“aware*” OR “interven*” OR “nudge*” OR “warn*” “protect*” OR “security indicators” OR “alert”)	Scopus
Grey	Phishing education/ training/awareness	“phishing” AND (“education” OR “training” OR “awareness”)	Google

Table 5
The inclusion and exclusion criteria.

Inclusion criteria	Exclusion criteria
I1. Studies the main focus of which is on phishing (defined in Table 1) education, training, and awareness.	E1. Studies that focus on automated phishing solutions (as defined in Jampen et al. (2020)) to counter phishing.
I2. Articles containing information about the research questions of this MLR.	E2. The study is a literature survey or review.
	E3. Full text of the study is not available.
	E4. The study is a short paper of fewer than 6 pages.
	E5. Studies not written in English.
	E6. Studies that have CORE rankings less than B.
	E7. Studies with CORE rankings of B and published before 2012.

on ways end-users can protect themselves from phishing (e.g., Sharma (2021)), which were not relevant to our study’s objective of examining the challenges and critical success factors of diverse anti-phishing interventions. Hence, following the guidelines of Garousi et al. (2019), we decided to use a separate approach for coming up with the search string for Google. To come up with the search string for Google, we explored grey literature from our pilot study to identify relevant terms that discuss anti-phishing interventions. Our search keywords for both academic and grey literature are shown in Table 4.

3.2. Study selection

A search was performed on the 2nd of May 2022 on Scopus, which returned 2760 articles of academic literature. During the execution of these searches, we did not restrict the searches by using any filter (e.g., time limit, type of publication, publication venue) to ensure the comprehensive coverage and collection of the relevant articles. We then applied the inclusion and exclusion criteria, as detailed in Table 5, to remove the articles that are irrelevant to the scope of this study.

The studies selected for this study were chosen based on their primary focus on Phishing Education, Training, and Awareness (PETA), as defined in Table 1. Specifically, we included studies that examine how to improve users’ ability to combat phishing attacks or investigate strategies to help users detect phishing attempts (inclusion criterion I1). It should be noted that this differs from a body of research that investigates users’ susceptibility to phishing attacks. In our pilot study, we encountered grey studies that discuss PETA but lack relevant information regarding the challenges (RQ1) or critical success factors (RQ2) of PETA. Therefore, we established inclusion criterion I2. As one of the goals of this study is to contribute to enhancing the usability of PETA interventions, we excluded studies that discussed anti-phishing solutions that do not require user intervention or users cannot see or act upon (automated solution), as indicated by exclusion criterion E1.

To avoid including low-quality papers, we adopt a quality assessment approach based on publication venues. We identified the CORE ranking⁵ of each publication venue of our search results and excluded

papers with rankings below CORE B. The CORE ranking is a process for ranking the academic Journals and Conferences in Computer Science (and related areas) that incorporates both expert domain knowledge and empirical data. The CORE ranking committee performs a thorough analysis of various factors, such as the citation count of papers published in the venue, the extent of involvement of leading researchers, as determined by metrics such as the author’s h-index, the acceptance rates of the venues, as well as the expertise and engagement of the Program Chair (PC), as gauged by metrics such as the PC’s h-index to rank the journals and conferences. The ranking system is regularly updated and refined to remain responsive to the changing needs and trends of the academic community (CORE, 2021b,a).

The CORE ranking is considered to be relevant outside of the Australian academic context because it utilizes widely recognized data and methodology as mentioned above for evaluating the quality of academic Journals and Conferences. Furthermore, the international academic community has widely accepted and recognized the CORE ranking system, as CORE rankings being used as a benchmark for study selection in many existing systematic literature reviews published in leading Journals and conferences (e.g., Souza et al. (2019), Croft et al. (2022), Franz et al. (2021)).

Again, we have considered CORE ranking over popular database SJR⁶ as the CORE database offered ranking for both journals and conferences during the study selection stage of our study.

To achieve an integrated and comprehensive analysis of up-to-date research in PETA, we removed papers with CORE rank B published before 2012. The primary rationale was to eliminate outdated studies in order to report challenges (RQ1) and critical success factors (RQ2) that are pertinent to current intervention design practices. Nevertheless, despite being published before 2012, we retained five CORE A* and A-ranked studies. This decision was justified by the fact that these studies provided recommendations that are relevant and applicable in the present context, for example, our included CORE ranked A* study published before 2012 (Egelman et al., 2008) provides suggestions like “providing clear choices for better understanding” and “interrupting users’ primary task to draw attention” for browser phishing warnings.

⁵ <http://portal.core.edu.au/conf-ranks/>, <http://portal.core.edu.au/jnl-ranks/>

⁶ <https://www.scimagojr.com/journalrank.php>

Table 6
Data quality assessment checklist for grey literature.

Criteria	Questions	Score
Authority	Q1. Is the source from a reputable organization?	Yes - 1, partly - 0.5, no - 0
	Q2. Has the author published any other article in the area?	Yes - 1, partly - 0.5, no - 0
	Q3. Is the author an expert in the area?	Yes - 1, partly - 0.5, no - 0
Methodology	Q4. Does the source clearly state the aim?	Yes - 1, partly - 0.5, no - 0
	Q5. Is the source supported by credible references?	Yes - 1, partly - 0.5, no - 0
Date	Q6. Is the publication date clearly indicated?	Full date/year only/ month and year - 1, day and month only - 0.5, no date - 0
Novelty	Q7. Does the source support or oppose a current position?	Yes - 1, partly - 0.5, no - 0
Outlet type	Q8. What is the source's outlet type?	1st tier - 1, 2nd tier - 0.5, 3rd tier - 0

These suggestions are relevant in the present context regarding the browser type and versions (Egelman et al., 2008). We have not considered h-index and citation count for inclusion/exclusion like some existing studies (e.g., Croft et al. (2022)) as the calculation of both of these metrics relies on citation count. Consequently, it is difficult to assess the quality of recently published studies (e.g., those published within the last three years) based on these metrics (Sabir et al., 2021).

Our Google search on the 22nd of April 2022 resulted in 121,000,000 items of grey literature. A large number of search results was attributed to the method of Google's search algorithm, which searches for specified terms across all indexed pages (Cascavilla et al., 2021). Moreover, these results contain academic studies and duplicate websites. Based on a careful examination of the search results, it was observed that results from page 17 to a few pages onward were either irrelevant or repetitive (we investigated up to page 20). As Garousi et al. (2019) suggested to stop the search when no relevant or additional information is found, we decided to limit our analysis to grey literature from pages 1 to 16 of the Google search results. Duplicate information from the same source shown on different pages was considered redundant and excluded. Academic studies from Google search results and duplicated Scopus results were also excluded. We further crawled through each link included in the websites of these 16 pages to collect further relevant studies. After reading the title, objectives, and the full article contained in these 16 pages, we retrieved 37 items of grey literature from Google.

3.3. Article quality assessment

Within the context of software engineering systematic literature reviews, quality assessment is conducted mainly through (1) explicitly defining assessment criteria and extracting them from primary studies, or (2) establishing research questions or inclusion/exclusion criteria that address quality concerns (Da Silva et al., 2011). In our study, we followed a similar approach mentioned in the second category.

We defined a CORE ranking-based study selection in the inclusion/exclusion criteria and opted to conduct a quality assessment for academic studies. We contend that the CORE ranking-based study selection method facilitated the selection of high-quality papers for our study. This is primarily due to two reasons.

Firstly, as mentioned before, CORE ranking is a rigorous process that involves experienced committee members ranking conferences and journals based on several widely accepted evaluation metrics (e.g., citation counts of papers published in the venue, author's h-index, acceptance rates of the venues) (CORE, 2021b,a).

Secondly, the Reviewers and Program Committees (PCs) of these venues are composed of experts from both academia and industry who perform peer reviews based on several metrics (e.g., novelty, correctness, contributions with well-supported methodology, impact, reusability, practicability) to ensure the scientific validity of the results (Soneji et al., 2022). The significance of this aspect is particularly pertinent in the context of our study. Our research aims to make a contribution to the enhancement of current and future anti-phishing interventions by identifying challenges (RQ1) and critical success factors (RQ2) through the synthesis of relevant literature. For this reason, the

selected studies should adhere to a rigorous and established methodology to ensure that their findings are credible and robust. Furthermore, by adopting a CORE rank-based study selection method, we were able to circumvent the subjectivity involved in defining and scoring the quality assessment, as noted in a previous study (Da Silva et al., 2011).

In contrast, as grey literature is non-peer-reviewed, we adopted a rigorous quality assessment checklist suggested by Garousi et al. (2019). Table 6 shows the criteria and the corresponding scores for each criterion. We have selected eight assessment criteria related to the reputation of the published authority (Q1), author of the publication (Q2), authors' expertise (Q3), etc.

The responses related to the eight criteria were classified into three categories: "Yes", "Partly", and "No", with corresponding scores of 1, 0.5, and 0, respectively, as adapted from the previous studies (Disanayake et al., 2022; Jiang et al., 2021). In contrast to the scoring process recommended by Garousi et al. (2019), where a score of 1 is assigned to "Yes" and 0 to "No", we deviated from this approach due to the variability in our data. Specifically, for criteria Q6 and Q8, we encountered three different types of data (discussed in the following paragraphs) that necessitated distinct scores to facilitate an accurate cumulative assessment similar to the other criteria.

To test the validity of our scoring system and to fine-tune the scoring metrics, a pilot study was conducted on a randomly selected sample of 16 articles from the grey literature. During the application of Q6, discrepancies were observed in the availability of publication dates for the articles. Some articles lacked any date information, while others only provided the year or specific day and month details. Given that the year signifies the recency of an article, it was considered the most crucial component of the date. Consequently, a score of 1 was assigned to articles with a known year of publication, regardless of the presence of day or month information. On the other hand, a score of 0 was assigned when no date was available. Articles that only provided the day and month received a score of 0.5 in Q6.

Regarding Q8, aside from the outlet types (tier 1, tier 2, and tier 3) defined by Adams et al. (2017) and later refined by Garousi et al. (2019) in the guideline, three additional article types (research reports, case studies, and guides) were discovered during the search process. To address this, we employed our understanding and observations to map these articles to the corresponding outlet types as follows: research reports were categorized as tier 1, case studies were categorized as tier 2, and guides were categorized as tier 3.

From the pilot study, we also observed that every grey article in our pilot sample scored a minimum of 5. The substantial variations in scores are mainly attributable to variability in criteria Q5, Q6, and Q8. As most of the articles in our pilot sample scored at least 0.5 for Q5, Q6, or Q8, we have chosen a cut-off value of 5.5 to exclude low-quality articles at this quality assessment stage (demonstrated in Eq. (1)).

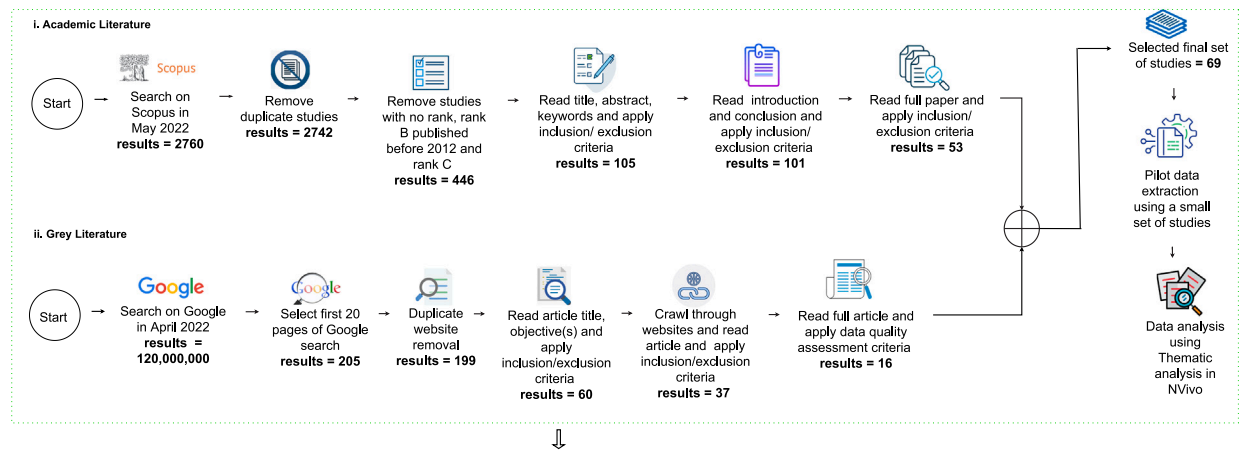
$$\text{Quality assessment score, } \sum_{i=1}^8 Q[i] > 5.5 \quad (1)$$

Here i is the question number mentioned in Table 6 and the value $Q[i]$ is either score 1, 0.5 or 0. Applying this process consistently across all items of grey literature resulted in the selection of 16 articles.

1. Planning the MLR



2. Conducting the MLR



3. Reporting the MLR

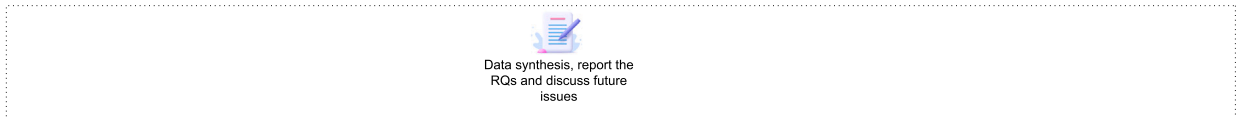


Fig. 1. Pictorial representation of our research methodology.

Our list of primary studies can be found in the supplementary online appendix.⁷

3.4. Data analysis

An overview of our data extraction and data synthesis process is provided in this section.

3.4.1. Data extraction

We systematically prepared and refined a data extraction form (please refer to our online appendix to see the data extraction form) to collect different types of data used for this study by following the existing guidelines (Kitchenham, 2004; Garousi and Felderer, 2017). Apart from collecting data for reporting our formulated research questions, we collected demographic information to gain a general understanding of the data included in this study (e.g., distribution and trends in the number of articles over time). We also collected recommendations for future research directions.

As suggested in the guideline (Kitchenham, 2004), data extraction was performed by more than one author (in our case, two). The first author extracted data from 49 studies and 14 grey studies, while the second author collected data from 4 studies and 2 grey studies. This collaborative data extraction approach provided us with the opportunity to include different perspectives by cross-verifying the data collected and developing a common understanding of the data that meets the required standards of quality and reliability. The data extraction form in Excel format was uploaded to the shared folder and discussed in the weekly research meetings among all the authors.

3.4.2. Data synthesis

The raw text collected from the primary studies, in accordance with the research questions guiding the investigation, was unstructured, encompassing a diverse range of information that can be challenging to interpret effectively. As a result, to gain a more comprehensive understanding of the significance and contextual relevance of the data, as well as to identify recurrent patterns that can facilitate the resolution of research questions, we conducted a thematic analysis. This method enables the exploration of the intricacies and subtleties of unstructured data by analyzing the raw text data in a systematic and rigorous manner (Dissanayake et al., 2022). Consequently, the use of thematic analysis allows for the elucidation of insights into the complexities of unstructured data.

We adopted the thematic analysis process discussed by Braun and Clarke (Braun and Clarke, 2006). Our thematic analysis process was conducted on NVivo, a tool for qualitative data analysis (Braun and Clarke, 2006). Our extracted data, stored in an Excel spreadsheet, was imported into NVivo, and then *open coding* was performed using this tool. *Open coding* is a process where labeled data (usually referred to as *code*) is obtained by breaking down the data into small components (Sbaraini et al., 2011) and labeling each component. We performed the open coding process through continuous iterations of extracted data (i.e., codes generated in the initial stage were modified and updated in later stages).

Initially, a pilot data extraction was performed by the first author with a set of five academic studies and two grey studies (randomly selected) to understand the pattern of the data. We scrutinized all the codes and grouped them into *themes* based on the similarities of the codes by utilizing the multi-layer structure of Nvivo. The first author revised the themes after weekly research discussions, and any suggestions or feedback from other authors were incorporated accordingly into the data analysis process. Fig. 1 demonstrates an overview of our research methodology and Fig. 2 displays an example of our data analysis process. The main findings are discussed in Sections 5 and 6.

⁷ https://drive.google.com/drive/folders/1IDH1BFPTJE73DH_zi9LR4TE_qJiYre?usp=sharing

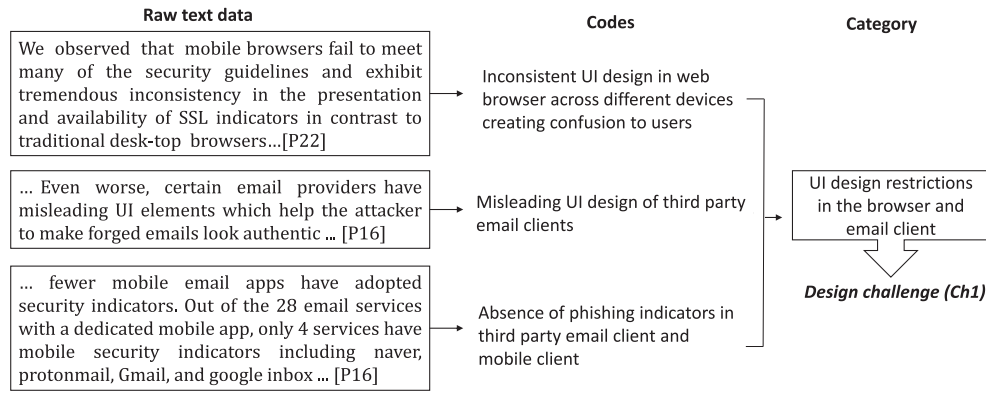


Fig. 2. An example of our data analysis process.



Fig. 3. Number of academic studies over years and CORE ranking.

4. Demographic description of selected studies

We report demographic data in this section in relation to PETA, such as the main publication venues for this area of research and the most/least investigated interventions. An overview of demographic data can help new researchers to gather useful information about the domain (Shahin et al., 2017). Some key insights about the demographic data are provided below:

- PETA first began drawing significant research attention in 2006. This domain has experienced rapidly increasing popularity in the last 4 years (2017–2021) (Fig. 3). Starting from 2006, our pool of primary studies includes no academic study from 2009 to 2011. Our pool of studies contains only A*-ranked publications from the year 2006 through 2008. Also, as we have decided to omit CORE B-ranked studies published before 2012 during our inclusion/exclusion stage, we do not have any B-ranked studies included before 2012. At the time of our search, we have not found any high-ranked (Rank A*, A or B) study published in 2022.

- Fig. 4 represents the venues of our selected academic studies ranging from journals, conferences, magazines, and workshop papers. International Conference on Human Factors in Computing Systems (CHI) is the most popular venue for publishing research in this field. Our pool of studies contains 22% studies (12 academic studies out of 53 academic studies) from this conference. Human-centric and security-related publication venues appear frequently; nevertheless, research outlets with other focuses, such as biomedical engineering and health sciences (e.g., Journal of the American Medical Informatics Association), also appear frequently, indicating that phishing is a common concern shared by researchers and practitioners from across different disciplinary fields.

- Fig. 5 demonstrates that we have more studies (both academic and grey) related to training, compared to education and awareness interventions. This data also indicates that education intervention is the

least explored category out of the three. The total number in Fig. 5 does not add up to 16 (number of grey studies) as one of the grey studies (P54) falls under both the phishing training and phishing awareness intervention categories. This dual categorization is attributable to the fact that the mentioned grey study contains information pertaining to both anti-phishing tools (which aligns with the phishing awareness category) and phishing training (which is associated with the training category).

- The type of grey studies we collected include whitepapers (tier-1, 43.75%), annual reports (tier-2, 25%) and blogs (tier-3, 31.25%) displayed in Fig. 6.

5. Findings of RQ1: Challenges in PETA

This section provides a holistic overview of the limitations and obstacles facing in the design, implementation, and evaluation stages of PETA. We use the symbol P[*] to refer to our primary studies.

5.1. Design challenges

Design limitations broadly cover limitations associated with the design and performance of anti-phishing interventions. The reported challenges mainly focused on the currently missing features in the design of phishing interventions.

5.1.1. Ch1. UI design restrictions in browsers and email clients

Design consistency in user interfaces is one of the critical usability attributes. A consistent design across different interfaces provides users with many benefits, including enabling them to transfer knowledge and skills across other similar systems, hence reducing their time and effort spent in learning to use the new systems [P22, P49]. A study documented that browser designers follow the same web security guidelines for designing user interfaces for both mobile and desktop browsers [P16]. This leads to *inconsistent UI design*, which creates confusion among users and increases their risk of exposure to phishing. The main reasons behind inconsistent designs are the lack of communication between mobile and desktop developers and the choice that developers need to make between usability and security. Mobile browsers have small display sizes; additionally, the padlock icon and the HTTPS URL prefix indicators in the address bar are hidden to accommodate the contents on a small display. Therefore, to make the content visible to the users and to keep a clean interface, less important information is often removed by the developers. Unfortunately, phishing indicators are one of the pieces of information removed from mobile displays in this process. Sometimes it is cumbersome for users to view the address bar in order to inspect the phishing indicators, which leaves phishing warnings unnoticed and exposes users to a greater risk of phishing attacks.

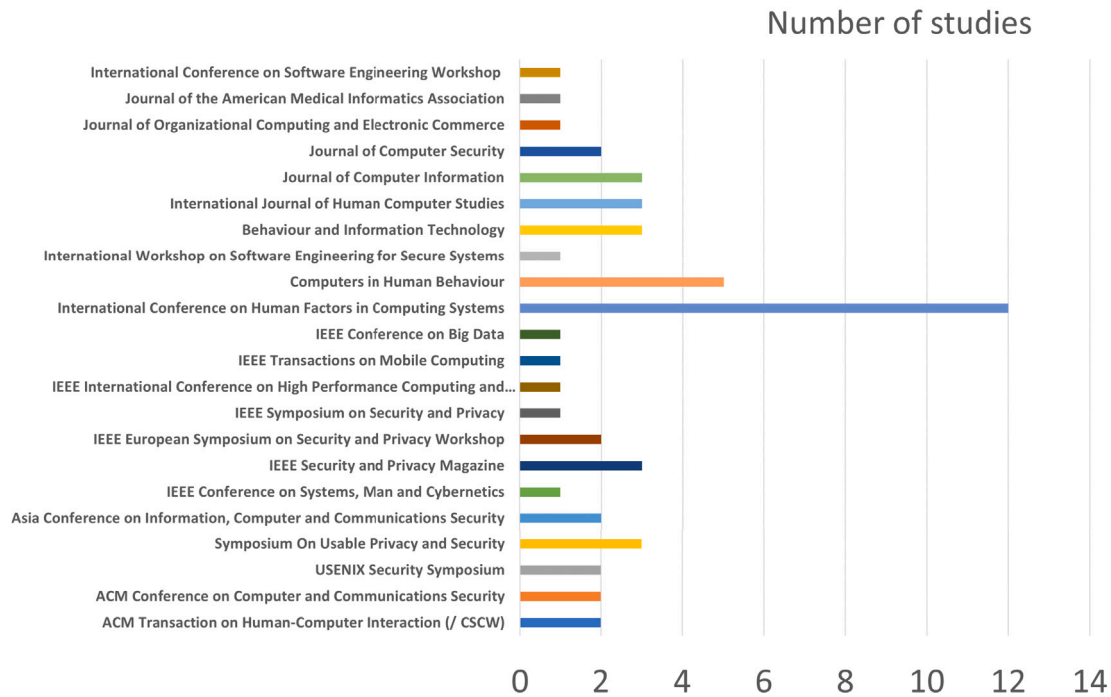


Fig. 4. Distribution of academic studies over type of venues.

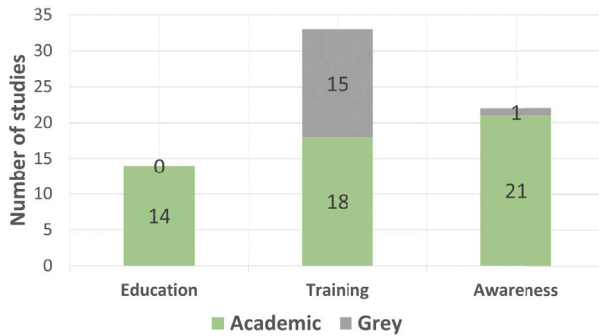


Fig. 5. Number of primary studies for each phishing intervention.

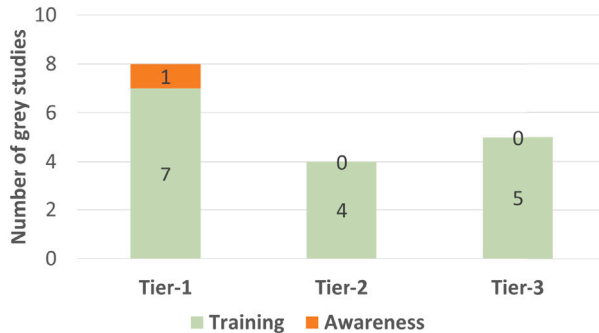


Fig. 6. Number of grey studies over tier types and phishing intervention.

Additionally, most mobile browsers do not use phishing indicators (e.g., extended validation SSL indicators). Many email providers and third-party email clients (Microsoft Outlook) do not provide warnings for forged emails. Moreover, some email providers (e.g., Gmail, Yahoo, Apple iCloud) incorporate *misleading UI design* which makes spoofed email appear legitimate, for example, by employing confusing colors to indicate the legitimacy of an email [P16]. This issue arises due to

miscommunication between email providers and end-users. Furthermore, the *absence of phishing indicators* grants potential attackers the opportunity to pass off the phishing email as a legitimate email.

5.1.2. Ch2. Content restrictions for phishing education and training

The effectiveness of phishing training largely depends on the content of the training material [P7]. Existing phishing training is *less engaging* as it allows one-way transmission of knowledge and does not provide immediate feedback to the users [P10, P19, P28]. Moreover, training materials often fail to capture the interest of users (e.g., a serious gamer will not be interested in playing educational games designed for casual gamers) [P36]. Some educational games have *complex interface and configurable scenarios* in the content, which renders them unsuitable for shorter training periods (e.g., CyberCIEGE) [P28]. Attackers use diverse attack vectors and a wide range of deceptions to lure users into disclosing their information. Unfortunately, existing training materials cover only *limited attack vectors* in the content. Consequently, users may only learn about some aspects of phishing attacks but not others, for instance, by learning about how to detect malicious URLs or deception cues, but remaining unaware of malware attachments that may also be enclosed in the email [P19, P24, P59]. Often training emails are *lengthy and wordy*, making it more time-consuming for users to make decisions. As a result, some users become confused about what messages the training email is trying to convey. This *time-consuming decision-making process* causes several problems to the users, including distraction and fearfulness, given that each user receives multiple emails per day [P5, P7]. Current designs of anti-phishing education materials do not consider users' *knowledge gap and misconceptions*, for instance, little consideration is given in the design of phishing training to human-centric factors such as users' state of mind (e.g., factors driving users to insecure behaviors and poor decisions making), how scammers operate, how users can be targeted, and users' strategies for dealing with phishing risks [P11, P19]. Contents containing *repetitive information* [P7] which users already have seen previously and the *presence of cultural bias* [P36] in the content (e.g., content on specific language, URLs from websites of a specific country) are also significant limitations in the design of current training contents. Repetitive training content does not add value to the knowledge and

consumes users' valuable time. Biased training content is only effective when users are familiar with the given information (e.g., websites or URLs from a particular country).

5.1.3. Ch3. Design constraints for anti-phishing warning UI interfaces

Several phishing indicators have been developed to warn users about phishing, including browser security toolbars such as Spoof-Guard (Teraguchi and Mitchell, 2004), Trustbar (Herzberg and Gbara, 2004), and site authentication image (a user-defined image selected for login, which enables users to verify the image before entering personal information) [P3]. Unfortunately, due to the *design similarity of phishing warnings with other security warnings* [P1, P28] and *lack of active interruption* [P1, P11, P14, P43, P44], phishing indicators often fail to attract user attention and lead to habituation to such warnings. A more salient function is needed to interrupt user actions in order to prevent users from succumbing to phishing attacks. Moreover, *frequent exposure to the warning* decreases users' neural activity, leading to carelessness, laziness, warning fatigue, habituation, and disturbance. As a result, users tend to automatically ignore phishing warnings [P4, P13, P14, P17, P18, P26]. *Unsuitable warning placement* is another reason why warnings go unnoticed. Most phishing indicators are small, making it difficult to attract user attention. For instance, in a phishing email where a malicious URL is the main hazard, current warnings are placed too far from the malicious links, which provides limited help to the users to identify the phishing URL. The warnings are not visually prominent and are not easily noticeable by the users [P2, P3, P5, P7, P11, P15, P25].

5.1.4. Ch4. Problems with anti-phishing warning content

Some email providers use warning banners to indicate phishing risks to the users. Such banner warnings typically do not provide sufficient explanation or reasoning to enable users to identify or assess phishing risk (for example, typically no explanation about why a link is deemed to be malicious is provided in the banner warning). The displayed information needs to be more comprehensive to enable users to make informed decisions and to enhance trust in the warnings. Although some current warnings offer a "learn more" button, they typically only contain general advice. The absence of specific information places an extra cognitive burden on users to locate the suspicious cues in the links. Instead of providing concrete information, current warnings are made unnecessarily *lengthy*, which is time-consuming to read [P41]. A *lack of comprehension* of the warning content, exacerbated by the absence of any justifications, leads to misunderstanding or ignorance [P14, P18, P25, P49]. Most often, the warning content is designed to target security-conscious and experienced users, ignoring the needs of novice or non-expert users. Additionally, the *lack of consistency in design practices of security warnings*, which vary significantly across different vendors, platforms, and browser versions, further creates confusion among users [P49].

5.1.5. Ch5. Performance limitations of anti-phishing tools

Anti-phishing indicators are designed to provide accurate information to users by accurately identifying potential sources of phishing. However, studies have shown that existing tools suffer from limitations in their usability and performance, which potentially impede the effectiveness of current anti-phishing tools in helping users detect phishing. The main source of usability issues arises from the tools providing inaccurate information to users by incorrectly identifying or failing to identify suspected phishing attempts (i.e. generating false positive or false negative results), which can further result in users' distrust of these tools. The current literature offers limited evidence to show that anti-phishing tools are indeed protecting users from phishing attacks. It remains unclear how, or to what extent, these tools assist users in determining a website's legitimacy. Designing usable phishing indicators remains an unresolved problem in the usable security domain [P1, P2,

P8]. Specifically, *False positive results* are an important limitation of existing anti-phishing tools. Some common reasons for falsely identifying legitimate sources of phishing attempts include: (1) delay in updating lists of known phishing sites, (2) improper maintenance of whitelist, (3) inaccuracy in performing successful detection (even the best phishing indicators miss 20 percent of phishing websites), (4) spyware infection in user device causing failures of site authentication image (e.g., Site-Key), which is used to protect user password during login to the page but cannot function in protecting users' personal information if users' computers are infected with spyware. Consequently, users learn to distrust and ignore the phishing warnings if previous warnings have mistakenly provided incorrect information (e.g., showing that a website is phishing when it is actually legitimate, displaying a phishing indicator even when there is no phishing risk, absence of phishing indicator when there is a high risk of phishing) [P1, P2, P3, P8, P10, P13, P14, P18, P24, P25, P28, P44, P49, P57, P69].

5.1.6. Ch6. Lack of attention to phishing indicators

For phishing indicators to serve their purpose, they must be heeded by end users. However, evidence shows an alarming frequency with which phishing warnings are ignored by end-users, due to a number of human-centric causes. First, users are more likely to ignore security warnings about possible phishing attempts during their online activities if they do not understand the risks and consequences of phishing attacks. This lack of knowledge results in a reduced likelihood for users to pay attention to phishing warnings [P1, P2, P3, P4, P8, P11, P14, P24, P28, P31, P36, P39, P44, P49]. This problem is compounded by the fact that many users misunderstand the nature of anti-phishing toolbars, as many users mistake the browser toolbar for an advertisement banner and are unsure if the toolbar is brought up by the browser or the website they are visiting. This lack of knowledge further makes it more difficult for users to interpret what a phishing warning is trying to convey. For example, prior evidence shows a lack of understanding by users of the nature and information conveyed by the anti-phishing toolbars such as the Neural-Information toolbar [P2]. Furthermore, a lack of knowledge about phishing can lead some users to apply the wrong anti-phishing strategy and distrust the security warnings (for instance, if users believe their approach to identifying phishing is correct, they would distrust any incongruous phishing warnings as wrong) [P2]. Another common reason why users fail to heed phishing warnings is when users have misplaced *confidence in the websites they visit* [P2, P8, P14, P24, P49]. For instance, users may cognitively deem a website to be legitimate due to its *look and feel*, causing users to ignore phishing warnings that show proof of authenticity of a website. Users also sometimes exclusively rely on one specific phishing indicator (e.g., site authentication image or SSL indicators) to the exclusion of others, causing users to ignore important information offered by other indicators which can contribute to determining the authenticity of a website [P3, P8]. Familiarity with a site or brand can induce users to trust the website [P11], thereby overlooking phishing warnings. For example, users tend to trust the website that they had previously visited. Moreover, reusing personalized phishing indicators allows attackers to develop an attack vector in the hope that users will use the same indicators for different applications [P31].

5.1.7. Ch7. Need to design specific training for spear phishing

Spear phishing is a phishing attack where attackers use victims' personal information to initiate the attack. Training users to detect spear phishing attacks is crucial as they are more effective and cause greater harm than regular phishing attacks, due to the level of personal relevance involved in phishing emails [P49]. However, the literature has identified numerous challenges that render it difficult to train users for spear phishing [P26]. Examples include emails mimicking standard business processes [P58], emails from known organizations [P1], timing of email receipt [P1], previous simulated phishing emails sent internally by organizations [P7], emails that do not request any

personal information [P7], messages from trusted sources (e.g., friends) [P15, P21, P44], and click-whirr response tendency (automatically responding to repeating events) [P44]. Due to the aforementioned reasons, users often ignore phishing warnings [P1] and face difficulty recognizing spear phishing emails [P7]. Often personal relevance invokes curiosity among users to click on a malicious link [P15]. The factors exacerbate users' vulnerability to spear phishing attempts, which will continue to be successful unless more attention is paid to the design of interventions aimed at training users to detect spear phishing attacks.

5.1.8. Ch8. Disregard for users' mental limitations during design

Human behaviors and decision-making are non-deterministic and unpredictable. A user who is security-conscious one day may act differently the next day, due to a number of human-centric factors such as illness or attention overload, resulting in greater exposure to security risk [P24]. Human-centric vulnerabilities constitute a greater risk factor than technical vulnerabilities to allow attackers to breach system security more easily [P24], as no complex cryptographic knowledge is required on the attackers' part to exploit human-centric vulnerabilities in phishing attacks [P49].

Despite the central role played by human perception and information processing in user decision-making, there remains a surprising lack of effort to incorporate users' cognitive constraints into the design of anti-phishing tools [P24]. Incorporating users' cognitive constraints into the design can be challenging. Unlike machines, human vulnerabilities cannot be patched with straightforward solutions, as human behaviors cannot be regulated, controlled, or changed by technical modifications [P15, P40].

Current anti-phishing technology fails to take into account some important human behavioral factors in the design process. For instance, (1) security is not the primary concern of users, whose attention is usually focused on other online tasks they are performing (such as reading books, checking email, or making online purchases), (2) users do not check security notification continuously [P1], (3) users cannot attend to everything at the same time due to cognitive limitations [P13], (4) when occupied with tasks that dominate their attention, individuals often fail to notice *highly conspicuous but unexpected* events [P13, P47], (5) reliance on *warning disruptions* can cause users to undertake a more passive view towards security consciousness and take no active actions to avoid phishing attacks [P13].

In summary, besides displaying phishing warnings to the users, there is a significant need to provide users with an *alternative options to complete their task* (e.g., suggestion for alternative website). Otherwise, warning disruption may cause some users to take risky actions to achieve their goals. Overall, the current lack of consideration for human-centric factors, such as users' cognitive constraints and lack of motivation or attention, in the design of anti-phishing tools and warnings can significantly restrict the effectiveness of these tools [P24].

5.2. Implementation challenges

Implementation challenges represent obstacles in adopting, deploying, and automating anti-phishing technologies, as well as weaknesses of the current policies and guidelines regarding anti-phishing defense.

5.2.1. Ch9. Anti-phishing technology deployment challenge

Existing studies have proposed small prototypes of anti-phishing tools or techniques (e.g., browser plug-ins) for the purpose of easy deployment on different platforms in real-world settings [P23]. However, real-world deployment of anti-phishing technologies faces several challenges [P6, P54, P57, P62, P65]:

First, organizations outsource phishing awareness and training material, however, it is often managed by internal staff members, which poses challenges to managing the materials optimally (optimal management is a shared responsibility). The involvement of external service

providers to provide support on phishing training program content or tool development makes it difficult to keep track of the changes and to measure the effects of phishing detection capability.

Second, it is difficult to make employees understand, identify, and safeguard their personal information due to a common misconception that ensuring security is only an IT problem rather than a responsibility shared by all personnel of an organization.

Third, distributed/siloed work environments and expanded infrastructure (e.g., new vendors, SaaS applications) create an enlarged attack surface, rendering it difficult for IT personnel to navigate as such navigation requires a team effort and coordination among team members.

Fourth, employees who work from home might lack sufficient infrastructure support. Moreover, the absence of adequate technical skills to set up a secure personal computing system increases their phishing risks.

Fifth, it is difficult to configure email clients to allow phishing training emails to be delivered to mailboxes, as training emails are often classified as spam by automated email security countermeasures [P28].

Finally, deploying anti-phishing tools and anti-phishing browser plug-ins can be a complicated process due to the interdependency of technological, organizational, individual, and procedural factors, such as browser platform dependency (e.g., Mozilla Firefox, Internet Explorer) and choices of IT standard or frameworks (e.g., Information Technology Infrastructure Library (ITIL)/Control of Business Objectives and Technology (COBIT)). Security service providers usually provide only high-level guidance, while ignoring the need for minor decision variations in specific organizations [P23, P38, P50].

5.2.2. Ch10. Technology adoption and usage challenges

Anti-phishing interventions (e.g., personalized security indicators, educational games, third-party anti-phishing tools, and anti-phishing training interventions) are developed to help users identify and assess phishing risks. Yet a number of usage-related challenges continue to impede widespread usage of these anti-phishing interventions and hamper their success. These challenges include: (1) some tools are difficult to use by non-expert users, (2) personal security indicators require extra efforts during installation and when using the applications [P37, P45], (3) the success of anti-phishing training depends on users' willingness and ability to learn and recall the information, and their capability to apply the learned information in subsequent situations [P19, P31, P45], (4) adoption of game-based education requires a level of prior knowledge and investment of time and efforts to build the software [P37, P45].

5.2.3. Ch11. Challenges due to complicated URL and domain name structures

Although users can be trained to improve their ability to identify malicious URLs, doing so is increasingly difficult due to the complex visual traps and textual manipulations employed by phishing attackers (e.g., hidden links, crafted texts, additional texts). Studies have shown that, even after training, it was difficult for users to detect small URL changes (e.g., swapping of two letters) [P45, P46, P47]. Complicated URL and domain name structures give rise to user confusion. Identifying well-concealed cues from visually inspecting URLs demands a great deal of attention and effort from users. Minor discrepancies can go unnoticed if user attention is distracted [P46]. This problem is exacerbated by the fact that some organizations lend their names for use by external parties: for instance, trustworthy websites, such as Microsoft, may provide hosting services for external web content (e.g., thereby enabling someone to pay Microsoft for web space and to create a website named *malicious.windows.net*). The resultant page is linked to a real Microsoft domain but its contents are controlled by the attackers.

5.2.4. Ch12. Obstacles to automating phishing incident response and anti-phishing training

Phishing attacks can cause damage within a few seconds [P50]. Therefore, it is essential to take prompt initiatives in response to phishing attacks. Organizations often run phishing simulations to train their employees or to test the ability of the employee to detect a phishing attack. Large phishing simulation campaigns can overload the help desk with phishing reports, which can impact the regular workflow of staff and hinder the effectiveness of phishing mitigation processes. Automating the ticketing system and phishing incident responses can help enable organizations to respond to phishing incidents promptly and save administrators time and effort. However, the task of automating incident responses is challenging because the performance of the incident response is mainly determined by the accuracy of the initial report, which in turn requires manual confirmation and validation by experts [P50]. Additionally, automating phishing training is also challenging as the content of a training email needs to be manually written and crafted by administrators to make it more realistic [P45].

5.2.5. Ch13. Exploitation of software vulnerabilities by attackers

Anti-phishing plug-ins (e.g., Anti-Phish, Spoofiguard) inhibit the transmission of user-sensitive information to the attackers' site by checking user inputs containing sensitive information. Nevertheless, such anti-phishing plug-ins are not foolproof as they are less effective when attackers use malicious JavaScript on their phishing websites. Such JavaScript provides attackers with opportunities to bypass monitoring phishing plug-ins. For instance, the use of JavaScript allows an attacker to listen to a critical press event on the client side and send each character back to the attacker's server before the user can press the submit button. Hence, users' sensitive information can be transferred to the attackers before the plug-in can detect that sensitive information. A solution to mitigate JavaScript attacks is deactivating JavaScript on web pages that include forms. However, it is not feasible to do so across the board, as many legitimate websites use JavaScript for form submission [P23]. Similar to a Javascript attack, attackers can also exploit other software-based vulnerabilities, such as by employing cross-site scripting (XSS) to inject malicious codes into the login pages executing on the client side to steal users' personal data. In this way, even an experienced user can be deceived into giving away personal information as the webpage containing malicious codes refers to a legitimate webpage [P49].

5.2.6. Ch14. Unguarded email clients and websites

Simple Mail Transfer Protocol (SMTP) does not use any built-in mechanism to prevent email spoofing. It relies on the SMTP extensions - Sender Policy Framework (SPF), Domain Key Identified Mail (DKIM), and Domain-based Message Authentication, Reporting and Conformance (DMARC) deployed voluntarily by the email providers for authentication. SPF helps senders maintain a DNS record containing a list of authorized IP addresses, which are allowed to send email from a specific domain. DKIM is a cryptographic signature used to sign an email to ensure that the email originates from an authorized source from that domain. In the absence of an SMTP authentication mechanism, attackers can send emails from spoofed email addresses. As the deployment of SPF and DKIM is not compulsory, few email providers adopted them. It is unclear how email providers handle an email that fails the authentication process [P16]. Furthermore, many e-commerce companies do not use SSL to protect their login page. SSL certificates ensure that users' data is protected and verify the authenticity of a website. Poor security practices by email clients and websites make it more difficult for users to identify the legitimacy of an email or website [P2].

5.2.7. Ch15. Limitations of current anti-phishing planning, policies, and guidelines

Studies have uncovered significant flaws in the existing anti-phishing guidelines, policies, and phishing training practices [P15, P42, P50]. Many organizations offer guidelines on their websites to enable their employees to learn how to identify phishing attacks. However, the information contained in such guidelines often contains generic information and does not include detailed explanations of the consequences of these attacks, nor information about potential advanced phishing techniques, such as clone phishing.⁸ Moreover, contradicting information included in the guidelines creates confusion and disturbance among users. Contradictory information also reduces readers' self-efficacy and ability to detect phishing [P42]. Readers of this information often misinterpret the information in the guidelines as a comprehensive list of possibilities, rather than mere examples. Consequently, users would recognize only the phishing cues mentioned in the guideline. Some information is incomplete or outdated, which can cause *security fatigue*. The absence of succinct and correct information also increases the likelihood for users to miss vital information or place trust in unreliable information [P15].

Organizations need to do more than merely adopt email security policies and guidelines in order to precipitate behavioral changes [P50]. More formal approaches are also needed to learn from past experiences involving previous phishing incidents. Poor planning in phishing training can diminish their impacts [P50]. For instance, if employees are summoned to the break room for phishing training on very short notice, this will likely reduce the effectiveness of the training session. Also, issues in the workflow process are not well understood to find the right tool or software, and often customized or outdated tools are selected without properly considering the best fit for phishing incident response [P50].

5.3. Evaluation challenges

The effectiveness of anti-phishing initiatives must be regularly evaluated to identify weaknesses and facilitate continuous improvement. This section discusses challenges encountered in evaluating the effectiveness and usability testing of PETA interventions.

5.3.1. Ch16. Lack of industry relevance in evaluation practices and settings

Many existing studies have methodological limitations which reduce the relevance of their findings to the industry. For instance, limited consideration of users of certain demographics (e.g., children) and the use of artificial settings (e.g., controlled environments) to observe user behavior induces sample bias, which in turn results in a lack of generalizability, as the findings and applications cannot be extended to other contexts [P1, P7, P13, P14, P18, P21, P26, P30, P32, P35].

For example, the findings of a study conducted with participants using Google Chrome might not be comparable with those using Mozilla Firefox, and the findings of a study conducting an experiment with university students or adults cannot be extrapolated to children or adolescents. Despite children's extremely high vulnerability to phishing risks due to their credulity and lack of experience, children are the single most overlooked demographic group in the literature [P21, P35], as most studies do not commonly take children into account.

Similarly, metrics used for evaluation also lead to erroneous calculations, for example, drawing the conclusion of phishing simulation based only on the click-through rate of simulation emails [P32].

5.3.2. Ch17. Complications regarding data collection and replicating user experience

A study setting that does not adequately replicate users' real-life phishing experiences would result in findings with low ecological

⁸ A type of phishing where a previously received genuine email is cloned to a malicious email (Chaudhry et al., 2016).

validity. Study participants might behave differently due to these environmental factors that are only present in the study setting [P14, P21, P40, P48]. Existing literature has reported several challenges regarding data collection and replicating users' real-life behavior.

For example, (1) collecting data from children is challenging due to the difficulty of obtaining and maintaining their attention during the study [P21], (2) participants may be disinclined to disclose truthful information about their past incidents of falling victim to phishing attacks, out of embarrassment or impression management [P40], (3) it is difficult to replicate users' real-life behavior during a phishing attack for various reasons: experimental settings lack the necessary element of risk [P3]; studies are often conducted with the help of a role-playing scenario [P3]; and Not evaluating user behavior in their regular working environment can affect their responses [P3, P43], (4) reimbursement or permission to opt-in before conducting the study poses some challenges.

5.3.3. Ch18. Insufficient usability and effectiveness evaluation of phishing interventions

Due to the continuous changes in the design of phishing interventions, it is essential to evaluate the usability of these interventions to understand user requirements and incorporate them into the design [P4, P8, P13, P17, P18, P30, P37, P40, P41]. A usability study has the potential to make significant contributions to the enhancement of the efficacy of anti-phishing interventions and the improvement of the user learning experience [P36]. Nevertheless, the expeditious pace at which updates are iteratively introduced to software components has the capacity to render previously conducted evaluative data outmoded. For instance, an investigation appraising the usability and efficacy of browser-based phishing warnings or anti-phishing toolbars becomes inapplicable upon the introduction of a new iteration of the browser, given the substantive alterations incorporated within the updated version [P4, P8].

The assessment of the effectiveness of methods or techniques employed for user education is crucial for enhancing their learning experiences. In some cases, organizations utilize techniques such as distributing leaflets to educate employees about phishing without conducting thorough evaluations to determine the practical efficacy of these approaches. The effectiveness of such methods might be diminished due to human-centric factors. For instance, individuals may disregard phishing instructions if they perceive them as irrelevant, especially if they lack prior experience with phishing and fail to recognize the relevance of these instructions to their own situations [P13].

Moreover, existing studies reported that rigorous empirical investigations using different methods and variables are required, such as the impact of individual and organizational factors on training effectiveness [P40], the role of different phishing cues in the decision-making process of phishing detection [P41], the impacts of different email types or contents during phishing attacks [P30], expert and non-experts' decision-making processes about phishing attacks [P41].

5.3.4. Ch19. Lack of sophisticated quantification of phishing training outcome

Organizations often run phishing training simulations to test their employee's ability to detect phishing attacks. This is done by sending employees fake phishing emails, and employees' ability is subsequently measured using different performance metrics. The most common performance metric used for evaluation is the number of times users end up clicking on phishing links. By counting the number of clicks, informs organizations about which employees are in need of access to anti-phishing instructions and training.

However, security tools and third-party software have bots that can click on all the links in an email in a sandbox environment. This process is executed to ensure no malicious URL is in the email. The URLs clicked by bots can be misinterpreted as originating from a human user, which

creates false positive results in the phishing simulation reporting. It can provide false insights about organizations' security [P55].

The measurement of phishing simulation outcomes can also be influenced by users' offline conduct such as prairie dogging, which refers to the phenomenon where an employee receives a phishing simulation email and lets other employees know about it. This practice can also affect click-on rates and consequently the measurement of outcomes from phishing simulations [P15, P59].

5.3.5. Ch20. Lack of post-training user knowledge retention practice

Although phishing training can improve users' knowledge regarding phishing attacks, such impact is subject to decay over time [P13, P21, P40, P45]. Studies have shown that phishing awareness returns to the pre-intervention level within a very short period [P21]. Even if users develop the ability to detect phishing through phishing training, they will soon struggle to remember the relevant information if they do not apply the learned information in practical situations [P45]. Currently, the effect and duration of this knowledge are seldom investigated, resulting in a limited understanding of how long the impact of PETA initiatives can persist and how often users require re-training [P7, P31, P34, P54].

Table 7 presents the congregated challenges along with the main key points. In order to differentiate the data derived from grey studies, the P numbers associated with such studies have been distinctly marked in green. Symbols E, T and A in the first column refer to the challenges related to phishing education, training, and awareness interventions respectively.

6. Findings of RQ2: Critical success factors in PETA

This section describes the data we collected to answer our second research question.

6.1. Design

This section discusses factors documented in the literature that improve the design of current anti-phishing interventions.

6.1.1. CSF1. Design of engaging and up-to-date training content

Leveraging situated learning in anti-phishing training can improve user engagement. Adopting situated learning helps prepare users for a heavy cognitive load associated with experiencing a real-world phishing threat by introducing them to a relatable simulation scenario. Presenting information in an interesting way (e.g., Gamification, interactive training modules and videos, including less text, more graphics in the content, or comic format) brings enjoyment to the users, develops user confidence, strengthens motivation, and helps enhance content consumption [P5, P10, P19, P28, P34, P36, P37, P61, P62].

The instructional materials ought to encompass diverse versions and variations to accommodate various learning preferences, recognizing that each person possesses unique learning styles and approaches. Furthermore, it is imperative to acknowledge the continuous evolution of phishing attacks, which underscores the potential for training content to swiftly become outdated. It is crucial to include recent cyber attacks and detailed information about how attackers operate and the types of tactics used by attackers [P57, P59].

6.1.2. CSF2. Design of comprehensible anti-phishing technology







Identifying a phishing email or URL is a complicated task. Many basic concepts need to be explained to end-users to enable them to gain and apply the knowledge when they encounter an email with a suspicious URL. A comprehensive report during this process can help users make an informed decision and satisfy their curiosity when they seek more explanations to improve their contextual knowledge. A well-explained report aims to provide recommendations or feedback along with the reasons for the recommendation. An explanation that logically

Table 7
Challenges in phishing education, training, and awareness.

Challenges	Key points (included papers)	#
Design		
Ch1. UI design restrictions in the browser and email client 	① Inconsistent UI design in web browser across different devices creating confusion to users [P22, P49] ② Misleading UI design of third party email clients [P16] ③ Absence of phishing indicators in third party email and mobile client [P16]	3
Ch2. Content restrictions for phishing education and training  	① Lack of engaging and interesting phishing education and training material [P10,P19,P28] ② Presence of complex interface and configuration in the game design [P28] ③ Repetitive training content [P7] ④ Disregard for user misunderstandings and interests [P11,P19] ⑤ Limited attack vector consideration [P19,P24,P59] ⑥ Disregard for both casual and serious gamers [P36] ⑦ Presence of cultural bias in the content [P36] ⑧ Time-consuming decision making process and lengthy training email [P5,P7]	9
Ch3. Design constraints for anti-phishing warning UI interfaces 	① Design similarity of phishing warnings with less serious security warnings [P1,P28] ② Frequent exposure causes warning fatigue [P4,P13,P14,P17,P18,P26] ③ Unsuitable warning placement [P2,P3,P5,P7,P11,P15,P25] ④ Absence of active user interruption [P1,P11,P14,P43,P44]	17
Ch4. Problems with Anti-Phishing warning content 	① Lack of comprehension and explainability [P14,P25,P49] ② Lengthy content [P41] ③ Distinct phishing warning design among vendors, platforms and web version [P49]	5
Ch5. Performance limitations of anti-phishing tools 	① Inadequate usability [P1,P2,P8] ② False positives and lack of reliability [P1,P2,P3,P8,P10,P13,P14,P18,P24,P25,P28,P44,P49,P57,P69]	15
Ch6. Lack of attention to phishing indicators   	① Ignorance due to lack of trust and understanding on phishing warning and training [P1,P2,P3,P4,P8,P11,P14,P24,P28,P31,P36,P39,P44,P49] ② Disregard to warning due to appealing web content and site reputation [P2,P8,P14,P24,P49]	14
Ch7. Need to design specific training for spear phishing 	① Difficulty to detect spear phishing due to personal relevance and familiarity [P1,P7,P14,P15,P21,P26,P49,P58]	8
Ch8. Disregard for users' mental limitations during design   	① Users' distraction by other tasks is not well considered [P2,P7,P8,P13,P14,P24,P47] ② Users' inattentiveness to phishing interventions have not been taken into account [P7,P13,P14,P17,P24,P58] ③ Current design practices unconditionally rely on user decision [P4,P15,P17,P24,P25,P40,P49] ④ No alternative options for users to help them complete their primary task [P2]	14
Implementation		
Ch9. Anti-phishing technology deployment challenge   	① Deployment difficulty of anti-phishing technologies due to interdependency on multiple factors and platform dependency [P23,P38,P50] ② Complicacy to safeguard employees in distributed and siloed settings due to enlarged attack surface [P6,P54,P57,P62,P65] ③ Training email spammed by email provider [P28]	9
Ch10. Technology adoption and usage challenges   	① Requirement of prior experience and investment in software for phishing games [P37,P45] ② Requirement of expertise and assistance from third-party services [P1,P8,P45] ③ Requirement of users' effort and willingness to use anti-phishing warnings [P19,P31,P45]	6
Ch11. Challenges due to complicated URL and domain name structures  	① Similar organization name in the URL [P2,P45] ② Difficulties to detect minor changes in URLs [P46] ③ User confusion to identify phishing website hosted by trustworthy websites [P45] ④ Presence of textual manipulations and complex visual tricks in the URL [P45,P47]	4
Ch12. Obstacles to automate phishing incident response and anti-phishing training   	① Handling phishing incident reports requires the need for human validation [P50] ② Embedded training deployment requires manual human effort [P45]	2
Ch13. Exploitation of software vulnerabilities by attackers 	① Use of malicious javascript codes by attackers to bypass monitoring phishing plugins [P23] ② Use of XSS by the attackers to inject malicious code into legitimate webpages [P49]	2
Ch14. Unguarded email clients and websites 	① Limited use of SSL indicator to protect website login page [P2] ② No built-in mechanism in SMTP to prevent phishing [P16]	2

(continued on next page)

Table 7 (continued).

Challenges	Key points (included papers)	#
Ch15. Limitations of current anti-phishing planning, policies and guidelines 	① Contradicting, incomplete and outdated anti-phishing recommendations in organizational websites [P15,P42] ② Choice of customized or outdated tools to manage IT incidents impact service quality and efficiency [P50] ③ Poor practice of training execution [P12,P59] ④ Lack of formal approach to gain experience from previous phishing incidents [P50] ⑤ Inadequate policies and guidelines to invoke user behavioral change [P50]	5
Ch16. Lack of industrial relevance in evaluation practices and settings 	① The neglect of young people to test and improve their phishing knowledge [P21,P35] ② Sample bias due to limited demographic consideration [P1,P13,P14,P30] ③ Failure to conduct usability testing in real-world settings [P1,P7,P26] ④ Poor evaluation practices results in unreliable outcome [P14,P18,P32]	10
Ch17. Complications regarding data collection and replicating user experience 	① Difficulty to emulate users real-life experience in phishing studies [P3,P43,P31] ② Ethical difficulties of conducting phishing studies [P48] ③ Challenges of phishing study due to bias induced by the participants [P14,P21,P40]	7
Ch18. Insufficient usability and effectiveness evaluation of phishing interventions 	① Negligible practical value and effectiveness evaluation [P4,P8,P13,P18,P37,P40] ② Inadequate empirical investigation on variables used in phishing training and detection [P30,P41] ③ Lack of understanding on user behavioral response towards phishing incidents [P17,P33,P41]	10
Ch19. Lack of sophisticated quantification of phishing training outcome 	① Difficulty in measuring user phishing training effectiveness due to presence of bots [P55] ② Impact of prairie dogging on phishing training program outcome [P15,P59]	3
Ch20. Lack of post-training user knowledge retention practice 	① Effectiveness of phishing interventions subject to dwindle over time [P13,P21,P40,P45] ② Lack of investigation on users' long term behavior change [P7,P31,P34,P54]	8

quantifies the decision made by the automated tool would increase users' trust in the systems. Users can also gain an understanding of the extent to which they can rely on the systems' decisions [P7, P8, P11, P14, P33, P39].

Offering anti-phishing recommendations to users in the form of visual examples and creating user-friendly URL patterns (e.g., using different colors for top-level domains and the rest of the URL) helps users to better absorb and retain the information [P8, P42].

In conventional phishing training or education methods, security experts decide what information to present to users. However, most often, security advice or help desk support is not available in real-time when the user experiences a phishing attack, during which there is little time to wait for expert advice or help desk support. Information from detailed reports can enhance the user's ability to comprehend the impacts of phishing attacks in order to provide appropriate responses to them in real-time. Providing an explainable report along with the automated anti-phishing technologies would encourage users to adhere to the warnings [P45].

6.1.3. CSF3. Diversity in training content to educate users on evolving phishing attacks

The diversity in learner behavior means that each user learns differently, therefore a cookie-cutter training method will only be effective for some users. It would be helpful to reach users in many different ways, for instance, by adopting various training methods such as flyers, posters, newsletters, and lunch and learn sessions [P58]. The continuously evolving nature of phishing attacks also demands changes and adjustments in the training content. A phishing game or training material containing only malicious domain and URL information leaves the user vulnerable to more advanced phishing attacks such as spear phishing. Leveraging diverse phishing attack vectors (link-based, data entry-based, or attachment-based) in the training template can enhance users' ability to detect a wider spectrum of diverse phishing attacks [P19, P61, P65].

6.1.4. CSF4. Consistency in training design

Consistency in training design helps users notice discrepancies, minimize confusion, and provide users with more opportunities to recognize inconsistent features. For example, a standardized template adopted by all the anti-phishing web pages reduces security fatigue. It also allows easier maintenance and updating. For instance, web designers can implement a tool more easily. To maintain consistency, researchers suggested a unified template for anti-phishing web pages proposed by a central agency such as CISA⁹ or ENISA¹⁰ [P42].

Studies also recommended that online services avoid using domain squatting techniques for domain names as it would be difficult for a user to identify the malicious domain if the legitimate domain uses domain squatting techniques, such as additional terms, unusual top-level domain (facebook.me), or subdomain (extra.facebook.com) [P46]. Using the same email styling in the organization is another recommended practice for design consistency [P41].

6.1.5. CSF5. Design of tailored phishing intervention

Poorly crafted and targeted phishing interventions will not be effective. For maximum reach and impact, phishing intervention should be appropriately personalized, for example, by including design of personalized training emails, adding local languages in the training content, designing realistic relevant template to train highly educated users, customizing the training style (where users can choose a preferred learning method), designing age-appropriate training tools (e.g., offering specific tools for children), dressing web application according to users' preferences, self-adaptive training where phishing simulation progresses in the level of difficulty based how well users perform, incorporating learning skills in the training design (e.g., consider casual and serious gamers in the phishing game design), customizing training content relevant to organizations and specific to job positions (e.g., managers or executives), and selecting training style suitable to

⁹ <https://www.cisa.gov/>

¹⁰ <https://www.enisa.europa.eu/>

the organization settings (e.g., use of text-based training instead of comics) [P7, P16, P21, P26, P35, P36, P40, P48, P49, P52, P53, P57, P58, P59, P61, P62, P63, P64, P66, P67].

6.1.6. CSF6. Improving the UI design

Studies have suggested that UI designers should devote greater attention to designing effective UI for email clients and phishing interventions, in order to draw users' attention, remove user confusion, and better support user queries.

For example, designers should focus on ensuring design consistency in the UI interface across mobile and web applications, to help reduce the phishing exposure of mobile app users. Other design techniques include deploying a more visually salient interface with noticeable color variations, removing misleading UI phishing indicators for unverified emails, and adding a support icon in the email client and intervention UI design to support user investigation (e.g., adding a *help me troubleshoot button*) [P4, P5, P7, P16, P51]. End-users are encouraged to avoid using the same personalized indicators for different interfaces [P31].

6.1.7. CSF7. Design of informative and concise warning

Anti-phishing advice alone is not enough to modify user behavior and reduce their phishing exposure. Abstract information presented to users should be coupled with concrete examples to achieve more effective communication and information retention.

Studies have shown that brief interventions have a relatively large positive impact. Too much information in the phishing intervention is unappealing, as inexperienced users may require an excessive amount of time to read and digest the information, creating an information overload. At the same time, interventions should be concise yet informative for educated and experienced users [P1, P5, P13, P18, P41].

Interventions should be designed in such a way that does not require lengthy decision-making from users (e.g., hovering over a link in every email received) to save their valuable time. Studies have shown that many users do not click on explanatory buttons such as *Learn more* or *More information*. Therefore, warning designers should not hide critical information and should not require scrolling down or additional clicks before such information is revealed to users. If a *Learn more* button is deployed, it should contain very detailed information to satisfy user curiosity [P4, P5, P25].

When users encounter a warning, a clear choice or advice should be provided on how to proceed. Simply asking users not to proceed might be counterproductive. An alternative path can be provided to them to finish their task [P1, P2, P5, P14].

6.1.8. CSF8. Incorporating users' psychological and behavioral aspects in the design

A phishing attack takes advantage of users' cognitive limitations in order to succeed. Therefore, it is important to take into account the limitations in human cognition, user misconceptions (how attackers operate), user assumptions, decision-making process (e.g., what specific cues users look for and how users interpret them), self-efficacy, and perceived threat in the design of phishing intervention [P9, P11, P18, P24].

Rather than informing users about the potential phishing risks, it is important to equip users to verify and assess the risks correctly. To design effective user-friendly phishing interventions, it is important to perform usability testing of phishing interventions to integrate user feedback in the design. Usability investigation allows the analysis of outcomes and helps identify trends over time to avoid repeating the same mistakes in the design [P22, P57, P61, P66, P67].

6.1.9. CSF9. Integrating phishing simulation with embedded training to facilitate education on demand

Organizations perform phishing simulations to test their employee's susceptibility to phishing. According to the existing PETA studies, coupling phishing simulation with training provides an effective approach to delivering an anti-phishing campaign.

Studies have suggested that receiving immediate instruction after a user clicks on a phishing simulation link can guide users on diverse phishing tactics without being involved in an actual phishing attack. Accompanying phishing simulations with learning units helps achieve desired behavioral changes. Studies found that embedding phishing simulation with training increases the reporting of the actual phishing emails. Studies have shown that a substantial time lag between the cause (e.g., clicking on the phishing link) and the effect (e.g., getting a phishing warning message about the email) may confuse users about why they are receiving the subsequent warning message, as the time lag makes it more difficult for users to identify the original click which triggered the warning message. Instead of scaring and confusing users, the learning content encourages them to be more careful and attentive when they next encounter a phishing email. By offering training intervention to users straight away after they make a mistake, users will be more appreciative of the *education on demand* [P5, P7, P12, P27, P53, P57, P58, P59, P67, P68, P69].

6.1.10. CSF10. Focus on active warning designs

A warning should be designed without expecting the users to keep security in mind while they perform their regular online activities. Phishing concerns should be integrated into the critical path of users' primary tasks to force users to deal with any warning before proceeding, as this helps users shift their attention from their regular tasks to the phishing warning. Due to habituation, users are less likely to read phishing warnings in their entirety. Sometimes users ignore the passive warnings due to the design similarity with less serious warnings. Therefore, to increase user willingness to read the phishing warnings, the warning should be designed differently from other trivial warnings by employing design features such as varying text size, color, highlighting, distorting the visual appearance of a phishing website, and placing the warning close to the suspicious link [P1, P2, P14, P20, P22].

In phishing education and training, users are often asked to hover their cursors over a link to check the legitimacy of the link. Users may accidentally click on the suspicious link while hovering over it. To reduce the risk when users hover over a link in a phishing email, researchers have suggested integrating action-based inhibitors (including a clickable link with a pruned URL) and adding a time delay before the link is clickable to allow users to overcome a small cognitive burden when dealing with the warning [P22, P25].

Adding interactive images on the login page, including *no-working links* on the site, is another possible forcing function that websites can adopt to prevent their users from submitting their credentials [P43, P44]. Adding an effective and non-obtrusive indicator on the site to notify users when they move from one domain to another (i.e., when clicking on a link leads them to an external website) provides another way of drawing user attention before they are exposed to potential suspicious URLs [P8].

6.2. Implementation

This section describes the recommendations provided in the state-of-the-art research and industry practices to enhance the execution process of anti-phishing approaches. We delve into a detailed analysis of these recommended strategies, shedding light on their operational intricacies and their potential to bolster the effectiveness of anti-phishing initiatives.

6.2.1. CSF11. Bringing key stakeholders on board to educate and encourage employees

In order to conduct a comprehensive and coordinated campaign of phishing simulation and training, C-suite executive officers are required to play a key role in its planning and implementation. For instance, extensive briefings before and during the phishing simulation and training are needed to better communicate with employees and resolve sensitive issues that may be encountered by individuals. Without such communication and sensitivity, a sudden and unexpected phishing simulation campaign may be criticized for undermining employees' self-esteem, targeting certain employees, or discriminating against specific groups. Managerial support from the C-suite helps employees understand that phishing security needs to be taken seriously. A group of champions must influence their peers about the necessity of phishing education and training and engage them in the process [P38, P40, P56, P57, P59, P61, P67, P68, P69]. University professionals and IT practitioners can also come forward to educate people about phishing. In this regard, some existing organizations such as (ISC)²¹ help to empower professionals on every aspect of cyber security [P21].

Another good strategy is to leverage external service holders' capability to take advantage of their specialization. In this case, external vendors must understand particular organizations' cultures, requirements, and goals. Internal and external expertise can be blended to validate employees' behavioral change after knowledge-based assessment, develop phishing awareness materials, manage and track user engagement in phishing training, and fully manage every aspect from top to bottom of a phishing simulation and training program [P54, P60].

6.2.2. CSF12. Strengthen authentication and encryption mechanisms in browsers and email clients

Studies have recommended improving the existing authentication and encryption mechanisms in browsers and email clients to create a strong line of defense against phishing attacks. Examples include the use of an SSL indicator for protecting webpages, the use of a single domain name by companies to prevent users from becoming confused by multiple domain names or IP addresses, getting SSL certificates verified by a trusted CA [P2], deploying browser-based user authentication to draw user attention, improving server-side and end-user based security protection, adoption of SMTP security extensions (such as SPF, DMARC, and DKIM) by the email clients to authenticate incoming emails, use of security indicator to alert users when unverified emails reach user inbox [P8, P16], and developing anti-phishing tools or apps that can deactivate JavaScript when the focus is on an input field of a submission form and reactivate it when the focus is not on the form, in order to reduce the keystroke monitoring performed by the attackers to launch timing attack [P16, P22].

6.2.3. CSF13. Feedback, reminders, and reinforcement to maintain phishing awareness among users

Along with training and testing, research has shown that providing friendly reminders and helpful feedback throughout the intermediate micro-training facilitates positive behavioral changes among users. Reaching out to the users about intermediate results, early communication, and providing beneficial feedback can enable users to assess what they have learned and to improve on an individual level. Frequent notifications and reminders may result in information overload. Therefore it is recommended to send notifications only when security violations have occurred [P53, P58, P60, P61, P62, P69]. Rewarding users' positive behaviors, such as by providing certificates, positive reinforcement during group meetings, gift cards, or increments of time off, is also helpful in motivating users to act in security-conscious ways [P30, P61, P66].

6.2.4. CSF14. Conduct GDPR-compliant and anonymous training to protect user privacy and avoid false training outcome estimation

To be effective, phishing simulation emails are often enriched with users' personal information to train users in preparation for a spear phishing attack. Such personal information needs to be incorporated appropriately, should be GDPR compliant (for organizations that provide service to EU customers), and should be safe from a data protection perspective. Using straightforward phishing simulation emails, which are easy to recognize, might reduce user motivation by making users feel overconfident about their preparedness for such attacks. In contrast, employing too sophisticated phishing simulation emails might make users feel deceived and tricked. Therefore, a balanced mixture of easy and challenging phishing simulation emails should be used [P29, P69].

Keeping the simulation results anonymous and general, conducting phishing training as a whole rather than targeting individual employees, and formulating the results in a way understandable to the users are among the recommended good practices [P59]. Finger-pointing of individuals who have failed a phishing simulation test can hinder their willingness to learn about phishing. For phishing simulations to be effective and sustainable, their primary emphasis should be on learning and maintaining anonymity (avoid collecting individual user's behavioral data). Individuals should not feel monitored and should be allowed to complete the phishing simulation at their own pace. Anonymous phishing simulation can also reduce the effect of prairie dogging. If prairie dogging occurs, even a zero-click rate after the phishing simulation would not be surprising. Prairie dogging distorts the phishing simulation outcomes and provides a false sense of an organization's preparedness for and susceptibility to phishing attacks [P59, P61, P62, P69].

6.2.5. CSF15. Providing phishing education and training to critical demographic group

Research has shown that highly educated IT professionals are no less vulnerable to phishing attacks than non-experts. This indicates that phishing education and training are necessary for everyone. However, when there are resource limitations, training priority should be given to groups that are more vulnerable to phishing attempts (e.g., employees who have access to the shared network), less motivated, and more careless [P13, P40, P53, P58, P60].

Surveys with children and teenagers found that they need support to deal with phishing attacks. Parents must be well-experienced or sufficiently informed to teach their children about phishing. Therefore, it is recommended that anti-phishing education should be included in school curricula. This gives rise to another significant issue of educating educators. Incorporating phishing training and education into mainstream education requires training teachers to make them feel comfortable and confident to deliver lectures on phishing in their classrooms [P21, P35].

Phishing awareness campaigns should also consider educating retailers. Retailers should be trained to provide reliable and faithful trust signals in their website design. This will help customers distinguish between legitimate and fake websites and maintain the retailers' business reputation [P64].

6.2.6. CSF16. Automating the phishing training to support the organization's security teams

Automation of the creation and delivery of training content, as well as phishing incident response, management, and reporting, can assist IT security teams in several important ways, including saving resources, assisting organizations to stay on the right tracks, minimizing efforts on installation and maintenance, allowing rapid response, limiting the organizational damage, reducing the number of victims, and making the help desk perform more efficiently while maintaining good practices. Automation support at the help desk can accelerate critical assessments needed to determine whether phishing reports

²¹ <https://www.isc2.org/>

are genuine and consequently enable prompt responses to phishing attacks, especially given their complexity and scale. Full automation of phishing incident management and reporting might require much work to be achieved. Therefore, automation is recommended for managing complex tasks when manual processes are too costly, which outweighs the cost of initial investment of installing such automated systems [P63, P67, P50].

Automation helps deliver personalized, frequent, and relevant training and helps with automatic threat identification and classification. Manually categorizing and personalizing the training content requires security teams to put in much effort and planning [P61].

6.2.7. CSF17. Better planning, policy management, and documentation on phishing training

Building sustainable phishing defense requires better policy-making and improved management. For instance, to enable simulated phishing emails to reach user mailboxes, IT systems require specific modifications. Otherwise, technical solutions deployed in the IT systems are likely to block the phishing simulation emails, assuming that they are harmful phishing emails. For example, these procedures include adding the email server's IP addresses to the technical IT system's whitelist and generating simulation emails that represent the entire spectrum of phishing simulations, in order to test whether such phishing emails can manage to get through the automated phishing detection systems to reach the user mailbox [P69].

Organizations need to conduct market research in order to select a vendor to provide phishing simulation and training services, which best fit the organization's requirements. The research process includes analyzing reviews of the service provider from unbiased sources (e.g., G2¹²) to assess vendor popularity, getting advice from peers, and browsing vendors' websites to gain more detailed information about the services that they provide [P61].

While planning for the phishing simulation, companies should prepare their help desk to support user investigations [P51]. The planning process also includes communicating transparently with the employees and notifying them about the purpose of the simulations to prevent discomfort [P30, P69].

Effective phishing simulations need to be supported and facilitated by well-structured planning documentation about the phishing simulation and training. The documentation should follow the policy guidelines and standards that describe important terms (e.g., phishing, spear phishing, smishing, vishing, URL), cover all training types and contents, execution details, and frequency (e.g., how the training would be conducted, how many times training would be performed), expected behavior of participants, and rewards and consequences [P26, P60].

6.2.8. CSF18. Enabling and encouraging individuals to report phishing

Phishing reporting is important for defending organizations against phishing attacks and measuring the effectiveness of phishing simulation and training interventions. Reporting allows users to actively participate in phishing defense and help build a security-conscious culture. The reporting data also enables the organization's security team to analyze suspicious emails that have managed to bypass the technical phishing defense system, which in turn allows the security team to update and strengthen their automated phishing defenses accordingly [P58, P63].

Establishing a phishing reporting system before conducting the phishing simulation allows users to contact the IT security teams as soon as they encounter phishing. Integrating an easy-to-use and in-client reporting button can reduce the ticket volume and burden on the help desk, as simulated emails are not directed to the help desk. Users can be encouraged to report when they learn about the button [P26, P50, P58, P63, P69]. Training users on when and how to report and explaining to users the positive impact of reporting phishing can make a difference in establishing a solid line of defense against phishing attacks [P58].

6.2.9. CSF19. Invest in both technical and socio-organizational functions and capabilities

In light of the increasing proliferation of human applications and devices, it has become imperative to mitigate human-related risk factors. Recent studies have revealed a notable shift in the strategies employed by cyber attackers, moving away from overcoming technology-based safeguards to exploiting vulnerabilities in end-users. Technology-based solutions, software updates, security patches, and firewalls sometimes fail to provide accurate detection (e.g., false negatives), leaving much of the responsibility of phishing detection to end-users. Hence, effective safeguarding against phishing attacks necessitates the integration of technological solutions with user-centered defense mechanisms. Integrating technology-driven solutions with user-centric defense mechanisms can mitigate users' excessive reliance on technical solutions [P3, P5, P12, P17, P26, P27, P28, P38, P41, P51, P53, P57, P58, P59].

To ensure better security and protection against phishing, a system that combines the strengths of two different detection approaches operating on different principles (e.g., combining blacklist- and whitelist-based phishing detection applications) can be deployed. This approach can take advantage of both applications, and one application could potentially detect a phishing email missed by the other [P18, P51].

6.3. Evaluation

This section reports the suggestions collected from the primary studies on critical success factors that contribute to improving the evaluation of the effectiveness and usability of anti-phishing interventions.

6.3.1. CSF20. Conduct intermittent short-time progressive training to reinforce users' phishing awareness

The benefits of engaging in phishing simulation and training are subject to knowledge decay over time. Participants' knowledge about phishing after 6–8 months of training is similar to their pre-training level. Compliance frameworks like ISO 27001 and GDPR demand continuous employee training on cyber security topics, including social engineering attacks such as phishing, to develop a strong human-oriented line of defense. Due to the knowledge-waning effect, to achieve desired behavioral changes, short training sessions, brief reminders, monthly phishing simulations with high-quality training materials delivered at the moment of failure (e.g., click on phishing link), and continuous training are recommended in the literature. Repetitive training is helpful for users who need help strengthening users' full understanding, which may be only partially attained during the first time of training. Studies have shown that phishing click rates are significantly reduced after repetitive training sessions [P5, P7, P27, P34, P62, P67, P69].

However, maintaining an appropriate balance of training frequency is important to avoid training fatigue [P52]. With proper planning, phishing simulation emails should only be deployed occasionally while targeting a specific optimal frequency (e.g., 4–6 simulations per year). Excessive training will create an extra burden on the organization as this involves gathering and analyzing a large number of statistics and reports [P56]. Progressive training (easy-to-hard training) on more challenging topics would systematically improve users' sensitivity to deception cues [P24]. Sending emails in a randomized order and not flooding every department with phishing simulation emails would be beneficial, as this will reduce the chance of employees discussing the simulation with their peers and also minimize the ticket load experienced by help desks [P24, P53, P56, P57, P68, P69].

Evaluation of employees' knowledge retention immediately after training (short-term retention) and weeks or months after training (long-term retention) are essential steps in the continuous evaluation process to obtain the desired training effect [P52].

¹² <https://www.g2.com/categories/security-awareness-training>

6.3.2. CSF21. Perform empirical testing and statistical analysis to improve and better support phishing training

Extensive empirical testing and evaluation of phishing simulation and training provide valuable data-driven insights, which can help organizations review their progress achieved by anti-phishing initiatives, optimize the training interventions, and plan long-term strategies to achieve more significant goals [P56, P57, P60, P61]. Setting a specific goal and establishing a baseline helps organizations stay focused on specific outcomes, recognize the specific changes required, track progress over time to ensure continuous improvement, and create a mature and robust phishing defense. Achieving cyber resilience takes time and requires patience and focused effort to assess users' shortcomings, in order to provide proper training [P54, P56, P58, P59, P60, P61, P68].

To improve the capability to gather empirical data, existing studies recommend the establishment of a governance data structure for users to report. Long-term impact assessment helps determine suitable training methods to increase user engagement [P31, P54, P57, P58].

Although regular, continuous training is recommended, regular training is costly and, in some cases, may be infeasible (e.g., training children in school regularly). Therefore, before conducting training, a rigorous evaluation of users' baseline knowledge should be performed. Moreover, challenging questions should be used in this initial evaluation to avoid the ceiling effect (minimizing the likelihood for participants to achieve the maximum test score in the initial evaluation), this will render future evaluation scores more informative as the increase or decrease of scores will be more visible [P21].

6.3.3. CSF22. Investigate if the phishing simulation is affected by false positives to avoid erroneous evaluation

Specific configurations (e.g., *presence of bots*) can cause false positives (e.g., high click rate) of phishing simulation assessment, leading to erroneous evaluation. Before conducting a phishing simulation, it is a good practice to check whether the current configurations will generate false positives.

For example, teams responsible for phishing simulations should check the inventory of all software, security solutions, and service environments and documentation, in order to identify whether they are performing any scanning, probing, or analysis. If the answer is yes, it is vital to deactivate these capabilities for certain IP addresses to avoid generating false positives during the phishing simulation. Also, identifying the *bot clicks* by checking the clicks made by web browsers and operating systems. If an organization uses any email security solution with an allow-listing feature, it is recommended to prevent phishing links from being scanned or clicked by bots by identifying the *bot clicks*.

To reduce false positives, participants should be informed to report phishing only through an approved reporting mechanism and avoid reporting phishing by using the email provider's default reporting button or function [P54]. Normalizing and re-scaling the click rates are also recommended to obtain a more accurate assessment of the outcome of phishing simulations [P32].

6.3.4. CSF23. Conduct user evaluation in their regular environment with realistic emails and measure delayed outcomes to replicate real-world settings

In real-world scenarios, users are not preoccupied with cyber security concerns when performing their regular online activities (e.g., checking email or conducting online shopping). Consequently, in a lab study, when users are asked to perform a security-related task (e.g., to identify phishing websites), users become more cautious, which disrupts their normal behaviors. Therefore, to produce generalizable results in a lab study, users' natural behaviors need to be preserved. While conducting phishing studies, researchers should observe participants' behaviors rather than interrupting; researchers should also obscure the purpose of the study by asking participants about other related subjects [P2, P4, P18, P43].

Evaluating user behaviors in their regular environment is an ideal approach, which can be achieved by employing the following recommended practices, in order to keep the experiment as realistic as possible and to achieve high ecological validity. Examples of these recommended practices include embedding simulation emails with users' regular email environments, asking users to install/deploy applications or browser extensions on their devices, collecting in-browser telemetry and training users with realistic emails [P4, P7, P31].

A study has suggested collecting real-time neural and eye gaze data by *brain-eye* measure to evaluate the reliability of a user's response to a phishing study. According to the authors, if the neural features of the users show that they could have been more attentive during the study, their response might not be valid [P17].

We have reported a total of 10 design, 9 implementation, and 4 evaluation-related critical success factors as demonstrated in Table 8. The P numbers in green symbolize the grey studies.

7. Insights from grey literature

The inclusion of practitioner perspectives and insights is critical for ensuring that phishing prevention strategies are effective and appropriately tailored to the evolving threat landscape. The grey literature in our study did not reveal any contradictory results compared to the academic literature. However, the grey literature was able to provide additional knowledge that would have remained undiscovered otherwise or reinforced and strengthened the results obtained from academic research. We further explain these points below.

- The omission of grey literature in our review can result in the loss of significant insights and information. Such a scenario may culminate in a disparity between the training administered and the practical threats that organizations confront, culminating in insufficient preparedness and heightened vulnerability to phishing attacks. For instance, important subject matters for inclusion in the design of phishing training content, as documented in CSF1 and CSF3, effective training techniques and strategies, as noted in CSF5, and recommendations for phishing training and incident response automation, as reported in CSF16, may be overlooked. Additionally, crucial aspects of training evaluation and knowledge assessment, along with suggestions for deploying and preparing IT systems for training, as highlighted in CSF17, might be excluded from consideration. Finally, valuable improvements to phishing reporting, as outlined in CSF18, may be missed, resulting in a lack of progress toward the enhancement of phishing defense mechanisms. Therefore, the inclusion of grey literature in our review is vital to ensure that all pertinent and relevant knowledge is captured, thereby enabling the development of comprehensive and robust phishing education, training, and awareness interventions.

- Our academic research findings are substantiated and reinforced by data gathered from grey literature. For instance, an academic study (P52) revealed that over-training should be avoided to reduce training fatigue. However, several other academic studies (e.g., P24, P53) recommended repetitive training as an effective strategy for reinforcing users' phishing knowledge acquired during initial training. This proposition was supported by multiple grey studies (e.g., P56, P57, P68) documented in CSF20. Thus, the inclusion of complementary evidence from grey literature sources serves to bolster and augment the results obtained from academic research.

- The incorporation of findings from grey literature in our literature review has potentially broadened and deepened the scope of our analysis and strengthened the credibility and validity of our conclusions. For instance, several academic studies (P14, P18, P32) have assessed end-users phishing knowledge using click-through rates (e.g., the number of times users click on a phishing link), a practice we identified as problematic in challenge Ch15. The reliance on click-through rates as the sole evaluation metric can yield misleading and inaccurate results due to the presence of bots, as noted in the grey literature (Ch19). Accordingly, the grey literature emphasizes the importance of carefully selecting appropriate evaluation metrics and relevant baselines to avoid obtaining erroneous outcomes that could create a false sense of security within an organization (CSF21).

Table 8

Critical success factors in phishing education, training, and awareness.

Critical success factors	Key points (included papers)	#
Design		
CSF1. Design of engaging and up-to-date training content E T	<ul style="list-style-type: none"> Incorporating situated learning to improve user engagement [P5,P10,P19,P28,P34,P36,P37,P61,P62] →Ch2.① Including up-to-date content in the phishing training [P57,P59] →Ch2.② 	11
CSF2. Design of comprehensible anti-phishing technology E A	<ul style="list-style-type: none"> Detailed report on anti-phishing efforts to persuade users to adhere to the warning and to support non-expert users [P33,P45] →Ch4.① Explicit anti-phishing protection tools to increase users trust on automated anti-phishing tools [P11,P39] →Ch5.② Integrate both visual and text example with explainability in the anti-phishing webpages [P42] Designing user friendly URL bar to remove users domain name confusion [P8] → Ch11.①,Ch.11② Providing users with reliable automated anti-phishing tools [P7,P8,P14,P33] → Ch5.② 	8
CSF3. Diversity in training content to educate users on evolving phishing attack T	<ul style="list-style-type: none"> Use of a variety of training content, a mix of tools for phishing training [P58] →Ch2.② Attack vector variation in the phishing training content [P19,P61,P65] → Ch2.② 	4
CSF4. Consistency in the design E A	<ul style="list-style-type: none"> Creating a standard unified template for anti-phishing webpages [P42] Organizations should practice using the same structure and features for legitimate emails [P41] Legitimate domain should avoid using common domain squatting techniques [P46] 	3
CSF5. Design of tailored phishing intervention E T A	<ul style="list-style-type: none"> Customized phishing training design for employees with power and authority in organization [P40] Prioritizing topics for training relevant to the organization [P16,P58] Taking account the target demographic into training design and execution [P48] Personalized training content [P26,P52,P53,P57,P59,P62,P66,P67] → Ch7.① Considering casual and serious gamers need in the game design [P36] Dynamic and self-adaptive phishing training [P63,P64,P66] Personalized communication style and medium for phishing training [P61,P62] Text training materials instead of comic materials in corporate settings [P7] Developing anti-phishing tools for children [P21,P35] → Ch16.① Web application dressing according to user preferences [P49] 	21
CSF6. Improving the UI design E A	<ul style="list-style-type: none"> Disabling misleading UI elements for unverified emails [P16] → Ch1.② Design of consistent phishing indicators for different interfaces [P16] → Ch1.① Use of various colors [P5,P7] Avoid using the same personalized indicators across different interfaces [P31] Adding a support button in the email client to support user investigations [P51] Adding an icon in email client indicating suspicious email [P7] → Ch1.② Limiting the number of warnings user encounters to reduce warning fatigue [P4] → Ch3.② 	6
CSF7. Design of informative and concise warning E A	<ul style="list-style-type: none"> Present abstract information using concrete examples [P1,P5,P13,P18,P41] → Ch4.② Incorporate progressive disclosure in the design [P4,P5,P25] → Ch2.② Warning should provide clear choice to the user [P1,P2,P5,P14] 	9
CSF8. Incorporating users' psychological and behavioral aspects in the design T A	<ul style="list-style-type: none"> Considering human vulnerabilities and decision-making process in the design [P9,P11,P18,P24] → Ch8.①,Ch8.②,Ch8.③ Perform usability testing to improve warning design [P22,P57,P61,P66,P67] → Ch5.① 	9
CSF9. Integrating phishing simulation with embedded training to facilitate education on demand T	<ul style="list-style-type: none"> Supplementing the phishing simulation with learning content [P5,P7,P12,P27,P53,P57,P58,P59,P67,P68,P69] 	11
CSF10. Focus on active warning designs A	<ul style="list-style-type: none"> Visual aids for safe browsing to draw user attention [P8] Link focused warning in the email client to grab user attention [P25] Warnings need to be actively interrupting users' primary tasks [P1,P2,P20,P22] → Ch3.② Design of phishing warnings should be different than trivial warnings [P1,P14] → Ch3.① Phishing indicators should distort the visual appearance of the website to help users distrust the phishing website [P1] Warnings should stay long enough to grab users' attention [P1] Action based inhibitor in the warning to reduce users cognitive burden and potential hazard of clicking malicious links [P22,P25] Use of forcing and negative training functions [P43,P44] 	9
Implementation		
CSF11. Bringing key stakeholders on board to educate and encourage employees T A	<ul style="list-style-type: none"> Important role should be played by the C-suite to secure the organization against phishing [P38,P40,P56,P57,P59,P61,P67,P68,P69] Universities and practitioners should come forward to educate people [P21] Leverage external service providers to support on phishing knowledge assessment and awareness material development [P54,P60] 	12

(continued on next page)

Table 8 (continued).

Critical success factors	Key points (included papers)	#
CSF12. Strengthen authentication and encryption mechanisms in browsers and email clients 	<ul style="list-style-type: none"> • Use single domain name and use SSL to encrypt websites [P2] → Ch14.① • Deploying browser-based authentication [P8] • Adoption of SMTP security extensions in email applications [P16] → Ch14.② • Deactivate or re-activate javascript to avoid keystroke or timing attack [P16,P22] → Ch13.①, Ch13.② 	4
CSF13. Feedback, reminders and reinforcement to maintain phishing awareness among users 	<ul style="list-style-type: none"> • Avoid frequent risk notification, avoid regular reminders, provide feedback to help maintain awareness [P53,P58,P60,P61,P62,P69] • Rewarding secure behavior [P30,P61,P66] 	8
CSF14. Conduct GDPR-compliant and anonymous training to protect user privacy and avoid false training outcome estimation 	<ul style="list-style-type: none"> • Conduct GDPR compliant phishing simulation [P26,P69] • Emphasizing the anonymity and learning aspect of the phishing simulation [P59,P69] → Ch17.② • Conduct random phishing simulation to reduce the effect of prairie dogging and estimate of organization's likelihood to fall victim to phishing [P61,P62] → Ch17.②,Ch19.② 	5
CSF15. Providing phishing education and training to critical demographic group  	<ul style="list-style-type: none"> • Raise retailers awareness about phishing along with their customers [P64] • Topics on anti-phishing training should be taught in the school to educate children [P21,P35] → Ch16.① • Everyone who has influence in organization's security should be trained [P53,P58,P60] • More focus on unmotivated and careless users [P40] • Teacher should be given priorities in terms of phishing education [P21,P35] → Ch16.① • Focus on vulnerable group for phishing education [P13] 	8
CSF16. Automating the phishing training to support organization's security teams 	<ul style="list-style-type: none"> • Automation in delivering personalized contents and automation in threat identification [P61] → Ch12.② • Automating phishing reporting and incident response processes with the use of improved tools [P50,P63,P67] → Ch12.① 	4
CSF17. Better planning, policy management, and documentation on phishing training  	<ul style="list-style-type: none"> • Improved phishing defense through improved management and policy making [P11,P38,P40,P50,P53,P54,P57,P67] → Ch15.⑥ • Structured and explainable policy and documentation of phishing training program [P26,P60] • Sending pre-notification to the participants to prevent discomfort [P30,P69] • Perform prior research and analyze the reviews on tools vendors [P61] → Ch15.② • Preparing IT system to avoid simulated email being filtered by technical filters [P69] → Ch9.② • Deploying post simulation help desk support to support users investigations [P51] 	14
CSF18. Enabling and encouraging individuals to report phishing  	<ul style="list-style-type: none"> • Establishing phishing reporting culture [P26,P50,P69] • Implementing easy-to-use, in-client phishing incident reporting tool [P58,P63] • Training users how to report phishing incident and explaining the benefits of reporting [P58,P60] 	6
CSF19. Invest in both technical and socio-organizational functions and capabilities   	<ul style="list-style-type: none"> • Effective phishing detection requires the combination of technological innovation and human intervention [P3,P5,P12,P17,P26,P27,P28,P38,P41,P51,P53,P57,P58,P59] • Combining strengths of multiple anti-phishing technologies [P18,P51] 	15
Evaluation		
CSF20. Conduct intermittent short time progressive training to re-inforce users' phishing awareness 	<ul style="list-style-type: none"> • Avoid over-training to reduce training fatigue [P52] • Multiple cycles of training to re-inforce phishing awareness [P24,P53,P56,P57,P68,P69] → Ch20.① • Repetitive training in a short time span [P5,P7,P27,P34,P62,P67,P69] → Ch20.① • Testing users' short-term and long-term knowledge retention after training [P52] → Ch20.② • Progressive training [P24] 	13
CSF21. Perform empirical testing and statistical analysis to improve and better support phishing training 	<ul style="list-style-type: none"> • An extensive test with challenging question to reduce repetitive training cost and avoid ceiling effect [P21] • Conducting phishing simulation [P56,P57,P60,P61] • Assessment of long term impact [P31,P54,P57,P58] • Selection of effective metrics and relevant baselines [P54,P56,P58,P59,P60,P61,P68] 	10
CSF22. Investigate if the phishing simulation is affected by false positives to avoid erroneous evaluation 	<ul style="list-style-type: none"> • Check if inventory management softwares are using any scanning, analysis or probing to identify unusually high volume of external IP addresses [P54] → Ch19.① • Normalize and re-scale click through rates for more accurate assessment [P32] 	2
CSF23. Conduct user evaluation in their regular environment with realistic emails and measure delayed outcome to replicate real world settings   	<ul style="list-style-type: none"> • Preserve users actual behavior to achieve results close to real world settings [P2,P18,P43] → Ch16.② • Use of field techniques for high ecological validity [P4] → Ch16.② • Testing users in their normal environment with instant corrective performance feedback [P7,P31] → Ch16.② • Realistic and equally difficult training emails to test the persistence of training outcome [P7] • Use of real-time brain-eye measure to collect transparent data [P17] → Ch17.② 	7

8. Threats to validity

In this section, we discuss the internal and external biases induced in different stages of research methodology and the strategies undertaken to minimize them.

- Although *study selection bias* due to the impracticability of collecting a large number of primary studies is an unavoidable limitation in systematic reviews (Dybå and Dingsøyr, 2008; Ampatzoglou et al., 2019), we endeavored to minimize the effect by systematically modifying our search string through a pilot study to capture all relevant academic studies.

- *Lack of generalizability* of the study outcome is another critical limitation common to all systematic reviews (Ampatzoglou et al., 2019). To ensure acceptable generalizability, we selected a popular digital library, Scopus, to collect our primary academic studies without restricting ourselves to publication year for high-quality venues. To collect our primary studies in the grey literature, we chose Google as a search engine and iterated through numerous pages of search results until new pages of results no longer provided relevant information that related to the subject of our study.

- To minimize the *conclusion validity* bias originating from different interpretations of the exact result (Ampatzoglou et al., 2019), the first author identified the codes and themes of the results, by applying an established method for analyzing and synthesizing qualitative data through thematic analysis, and shared the information with all other authors. Code books were updated based on the suggestions and feedback provided by other authors in weekly meetings.

- Another common threat to validity is *publication bias* arising due to researchers' tendency to report positive results compared to negative ones (Kitchenham and Charters, 2007). However, in our MLR, we reported adverse effects as challenges, thereby ameliorating potential publication bias.

Despite all the aforementioned mitigation strategies undertaken in this study, we acknowledge that our reported list of challenges and critical success factors may not be exhaustive due to the inevitable internal and external biases (e.g., missing primary studies, grey study quality assessment, usage of the non-comprehensive database, and thematic analysis process). Therefore, we encourage the readers to take this into consideration while reading our study as we believe that our study serves as a valuable starting point and one-stop-shop for readers to gain familiarity with the current state of practice in this domain, which can enable readers to explore possible areas that require further attention and investigation.

9. Discussion and future research directions

In this section, we summarize and discuss the findings to provide an overall understanding of the key outcomes of our study. In relation to RQ1, drawing upon the evidence from 69 primary studies from both academic and grey literature, we discovered 8 design challenges, 7 implementation challenges, and 5 evaluation challenges. With regard to RQ2 (critical success factors), we identified 10 design CSFs, 9 implementation CSFs, and 4 evaluation CSFs. We only summarize the highly reported challenges and critical success factors in the design, implementation, and evaluation stages of PETA.

- Our MLR uncovered the demand for improving the UI design of phishing warnings (a predominant design challenge discussed in 24% studies), specifically in relation to warning design variation, active interruption, warning placement, and warning exposure (Ch6). This evidence indicates the significant need to improve the design of user interface of anti-phishing warnings to make them more accessible to end-users.

- A major challenge in the implementation of anti-phishing technology (Ch9) occurs during its deployment (13% studies), when problems arise due to platform dependency and distributed work settings. This suggests that developers and practitioners test their prototypes on

different platforms before finalizing their models. This evidence also highlights the need to develop mechanisms that help organizations' security teams safeguard employees in a distributed office environment.

- Our findings disclose that the main evaluation challenges (discussed in 14% of studies) are limited industry relevance of the findings of studies (Ch16) and inadequate usability evaluation of phishing interventions (Ch18). This indicates the importance of *more rigorous* evaluation across a more significant array of demographic groups to test the usability of phishing interventions.

- In terms of design, the main critical success factor mentioned by 30% of studies is incorporating individual user needs into the design of phishing interventions (CSF5). This suggests the urgency of exploring the needs of individuals, based on their age, educational qualifications, geographic location, profession, physical disabilities, language preferences, and other idiosyncrasies, in order to improve their capabilities to detect phishing attacks.

- When it comes to implementation, most of our retrieved studies (21%) indicate the equal importance of adopting reliable technological anti-phishing solutions and human-oriented education and training. Studies emphasized combining technical solutions that operate on different principles with providing education and training to users to reduce their dependency on technology-based solutions.

- Regarding evaluation, most of our studies (18%) recommended conducting follow-up training sessions to test users' knowledge retention and to reinforce users' phishing knowledge.

- In Table 8, we map the critical success factors with corresponding reported challenges. From this mapping, it is evident that challenge Ch10. *Technology adoption and usage challenges* do not currently have any recommended success factor documented in the literature that can help overcome the challenge. This provides an opportunity for future researchers to investigate such gaps in the literature, for example, by examining how the anti-phishing tools and applications can be simplified, how third-party dependency of anti-phishing tools can be minimized, how users can be encouraged to install and use the anti-phishing applications, and how requirements of user skills and experience can be minimized for application installation and usage.

In the following sections, we discuss some open issues based on our findings and gaps in the literature identified in this MLR, which provide fruitful avenues of investigation for future researchers.

9.1. Equipping anti-phishing systems with explainable capability

In his study, Metaxas (2009) states that “*It is the users' right and responsibility to decide what is acceptable for them. Their browser, their window to the cyber world, should enhance their ability to make this decision.*” Our findings indicate that current anti-phishing tools are lacking in this respect, as they fall short of providing adequate explanations about specific phishing risks to users to enable users to make their own informed assessments (Ch4●). As a result, many users still fall prey to phishing even after receiving a warning due to a lack of understanding. This lack of understanding often leads to phishing warnings being ignored (Ch6●). Several studies in our study pool recommended explainable and comprehensive anti-phishing tools to motivate users to adhere to phishing warnings by providing them with a detailed level of understanding of the reasoning (CSF2). Therefore an anti-phishing tool or mechanism should provide users with context-related information for various phishing-related problems that they may encounter. For example, assessing a phishing URL requires a user to have better knowledge about the structures of the URL to make an informed decision. To help users make an informed decision, browser developers can make the URL bar more user-friendly [P8]. Whenever an anti-phishing tool detects any unauthorized signal, it should provide users with adequate explanations and information about the warning, which would enhance users' knowledge about phishing and enable them to understand and assess future phishing risks (Reuter et al., 2022). A lack of information results in a lack of trust on the part of users

and, consequently, users' under-reliance on anti-phishing tools (Schuetz et al., 2022). In this regard, the system can offer additional information which helps users to make a correct choice [P33]. Explaining an anti-phishing tool's reliability (e.g., how the tool detects phishing attack, its confidence level in its decision, and the logical consequences of a decision) also increase user trust and reliance on the anti-phishing tool [P33].

Practitioners who specialize in human-computer interactions can contribute to the design of anti-phishing tools that provide additional necessary information. When users encounter phishing attacks, they often seek suggestions from the help desk in making their decisions. Anti-phishing tools with analytical capability can reduce help desk traffic [P44]. Researchers and practitioners can borrow the concept of *explainability* (Doshi-Velez and Kim, 2017) from machine learning and artificial intelligence in designing anti-phishing tools to provide additional information about the detection process applied by these tools. Research has shown that explainability in the design helps users gain trust in the system (Dzindolet et al., 2003). Explainable anti-phishing tools can bring several advantages. For example, the logical reasoning about the underlying model would increase their justifiability and transparency, help debug certain flaws and improve the detection approach, and assist curious users in learning about the relationship and patterns employed to detect phishing attacks (Vilone and Longo, 2020).

9.2. Platform for realistic phishing security testing

Our MLR reveals that existing phishing studies have faced difficulties in replicating users' real-life experiences with phishing within the study setting. Due to privacy reasons, it is difficult for the researchers to collect personal data, such as users browsing history, that could be useful in understanding user susceptibility to phishing and, accordingly, designing phishing interventions to educate and train users [P4]. Research outcomes can be affected by participants not following instructions (e.g., children discussing the study with their peers) [P21]; similarly, other studies documented that users did not follow or read the instructions before answering the survey questions [P17]. Some studies collected user feedback in an online survey or by adopting a role-playing scenario. Role-playing scenarios negatively impact users' security consciousness, as users tend to behave in less secure manners in the absence of real-world threats [P3]. Moreover, due to ethical restrictions, some studies reported the vulnerability to phishing by measuring only click rates. Ethical constraints also impose restrictions on how studies can be conducted, for example, users should not be tricked into giving away their personal information during the study [P16]. However, clicking is not the final stage in an actual phishing attack. Instead, the next step in a phishing attack involves users revealing personal information, which cannot be replicated ethically in a study setting. Game-based education typically trains users in an artificial environment which is dissimilar to the natural setting where phishing occurs [P26].

Some commercial phishing simulation and training platforms, such as knowB4 and Hoxhunt, provide organizations the opportunity to test users' phishing susceptibility in a relatively realistic setup. These platforms provide mechanisms to embed phishing simulation emails with clients' regular emails. Then a phishing training page will pop up if users click on the phishing link (to instruct users on how to respond when they see similar emails in real life). However, the effectiveness of this phishing simulation platform is affected by the impact of prairie dogging (challenge Ch19.1), as bias is added when users know that their phishing knowledge will be tested [P21]. Therefore, to achieve the desired results, researchers and practitioners can focus on developing a platform that can provide a reliable mechanism close to users' real-life phishing experience to test phishing security. Researchers can take a conceptual idea from existing usability testing platforms such as Maze.¹³

9.3. Automated tool to assess users' attentiveness during online engagement

To improve the usability of the current phishing interventions, the literature suggests conducting usability evaluations and collecting users' feedback during their interactions with phishing interventions (e.g., CSF21). The design improvement primarily relies on users' feedback. Sometimes it is difficult to collect unbiased user feedback from surveys, as users try to hide negative experiences [P40]. Studies documented the need for a practical approach to identifying whether a participant has read the phishing intervention content [P7]. Designers of Phishing interventions can only improve their design if they receive reliable user feedback. We recommend that developers of anti-phishing tools should invest in the development of automated tools to identify users' attentiveness during online engagement, for example, eye-tracking and neuro-imaging devices to collect users' neural and eye gaze features in real time to identify users' state of alertness.

While it is acknowledged that certain studies have already ventured into a similar investigative approach [P17], it is imperative to note the necessity for further comprehensive research. Future investigations should be undertaken to validate and refine the proposed mechanism on a broader scale. By delving into a wider array of contexts and user demographics, researchers can ensure the robustness and applicability of the developed tools, ultimately enhancing the efficacy of anti-phishing measures in the digital landscape.

9.4. Adopting automated and adaptable individualized approaches

Our MLR broadly highlights that a *cookie cutter* approach to phishing education and training would not be adequate to cater for specific demographic groups across a wide spectrum. Aspects of phishing education and training, from training content to style (Reuter et al., 2022; Salamah et al., 2022; Bullee and Junger, 2020), should be personalized to meet individual user needs in order to be helpful and effective. For example, a story-based training style is more effective for children [P21], whereas comic-based training content is unsuitable for corporate settings [P7]. Organizations should also consider modifying the content of their training program according to the routine business emails their employees receive. However, a fundamental challenge would be overcoming the manual effort required to personalize the training content. Incorporating automation can be a promising solution to conserve the time and effort of cyber security teams by replacing the manual process of choosing and modifying content suited to specific users. We suggest developing an adaptable phishing training approach where the difficulty level of the training content will be adjusted based on user knowledge. This can be achieved by conducting an initial phishing test to understand users' knowledge level and then delivering the training content automatically according to users' skills and abilities. Automatically clustering users of similar knowledge can also be helpful in providing personalized training more efficiently. This will prevent individual users from becoming demotivated due to the training content not being pitched at their level of need.

10. Conclusion

In this study, we provide the first systematic overview of the challenges and critical success factors across the stages of design, implementation, and evaluation of PETA interventions by synthesizing and consolidating the empirical findings from prior academic and grey literature. This study not only offered novel insights to deepen our existing understanding of anti-phishing interventions but also articulated a set of valuable suggestions to enable industry practitioners and stakeholders to harness of power of prior research findings when designing, implementing, and evaluating their phishing interventions, in order to enhance the success and effectiveness of their endeavors to protect organizations from future phishing attacks. The identified challenges can help researchers and practitioners, particularly those new to

¹³ <https://maze.co/>

designing, implementing, or evaluating anti-phishing interventions to understand the ongoing potential obstacles and roadblocks that hinder their endeavors to achieve anti-phishing effectiveness. Additionally, our findings on critical success factors provide a set of guidelines to help researchers and practitioners improve the design, implementation, and evaluation of PETA interventions. Moreover, we provide a valuable mapping of the existing challenges to the critical success factors in tabulated form that can help mitigate each challenge. These insights, summarized in tabular format, are expected to provide significant assistance to researchers and practitioners to overcome the obstacles and to improve the effectiveness of their anti-phishing endeavors.

Our stage-specific critical success factors can be utilized by future researchers to develop stakeholder-specific guidelines targeting different practitioner groups. For example, our design-focused critical success factors can be further subdivided to generate guidelines for anti-phishing warning designers, domain engineers, tool developers, or browser designers. Similarly, our implementation and evaluation-focused guidelines can be tailored to inform security teams and senior executives in the organization. Achieving these stakeholder-oriented perspectives will enable practitioners easier access to relevant information and avoid wasting time and effort to wade through general information that is less relevant to a particular practitioner group.

CRedit authorship contribution statement

Orvula Sarker: Conceptualization, Data curation, Investigation, Validation, Methodology, Writing – original draft, Writing – review & editing. **Asangi Jayatilaka:** Conceptualization, Data curation, Validation, Methodology, Writing – review & editing. **Sherif Haggag:** Validation, Writing – review & editing. **Chelsea Liu:** Validation, Writing – review & editing. **M. Ali Babar:** Discussing and supporting the study idea and scope, Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

As the submitted work is a multi-vocal literature review, the data are already publicly available.

Acknowledgment

The work has been supported by the Cyber Security Research Centre Limited, Australia whose activities are partially funded by the Australian Government's Cooperative Research Centres Programme. The manuscript has been read and approved by all named authors.

References

- Adams, R.J., Smart, P., Huff, A.S., 2017. Shades of grey: guidelines for working with the grey literature in systematic reviews for management and organizational studies. *Int. J. Manag. Rev.* 19 (4), 432–454.
- Alani, M.M., Tawfik, H., 2022. PhishNot: A cloud-based machine-learning approach to phishing URL detection. *Comput. Netw.* 109407.
- Althobaiti, K., Jenkins, A.D., Vanica, K., 2021. A case study of phishing incident response in an educational organization. *Proc. ACM Hum.-Comput. Interact.* 5 (CSCW2), 1–32.
- Ampatzoglou, A., Bibi, S., Avgeriou, P., Verbeek, M., Chatzigeorgiou, A., 2019. Identifying, categorizing and mitigating threats to validity in software engineering secondary studies. *Inf. Softw. Technol.* 106, 201–230.
- APWG, 2022. Keeping up with Phishing. URL: <https://apwg.org/trendsreports/>.
- Arachchilage, N.A.G., Love, S., Beznosov, K., 2016. Phishing threat avoidance behaviour: An empirical investigation. *Comput. Hum. Behav.* 60, 185–197.

- Bada, M., Sasse, A.M., Nurse, J.R., 2019. Cyber security awareness campaigns: Why do they fail to change behaviour? *arXiv preprint arXiv:1901.02672*.
- Baki, S., Verma, R., 2021. Sixteen years of phishing user studies: What have we learned? *arXiv preprint arXiv:2109.04661*.
- Benzies, K.M., Premji, S., Hayden, K.A., Serrett, K., 2006. State-of-the-evidence reviews: advantages and challenges of including grey literature. *Worldviews Evidence-Based Nurs.* 3 (2), 55–61.
- Braun, V., Clarke, V., 2006. Using thematic analysis in psychology. *Qual. Res. Psychol.* 3 (2), 77–101.
- Bullee, J.-W., Junger, M., 2020. How effective are social engineering interventions? A meta-analysis. *Inf. Comput. Secur.*
- Burda, P., Chotza, T., Allodi, L., Zannone, N., 2020. Testing the effectiveness of tailored phishing techniques in industry and academia: a field experiment. In: *Proceedings of the 15th International Conference on Availability, Reliability and Security*. pp. 1–10.
- Butijn, B.-J., Tamburri, D.A., Heuvel, W.-J.v.d., 2020. Blockchains: a systematic multivocal literature review. *ACM Comput. Surv.* 53 (3), 1–37.
- Caballero, A., 2017. Security education, training, and awareness. In: *Computer and Information Security Handbook*. Elsevier, pp. 497–505.
- Cascavilla, G., Tamburri, D.A., Van Den Heuvel, W.-J., 2021. Cybercrime threat intelligence: A systematic multi-vocal literature review. *Comput. Secur.* 105, 102258.
- Chaudhry, J.A., Chaudhry, S.A., Rittenhouse, R.G., 2016. Phishing attacks and defenses. *Int. J. Secur. Appl.* 10 (1), 247–256.
- Chen, J., Mishler, S., Hu, B., Li, N., Proctor, R.W., 2018. The description-experience gap in the effect of warning reliability on user trust and performance in a phishing-detection context. *Int. J. Hum.-Comput. Stud.* 119, 35–47.
- CJ, G., Pandit, S., Vaddepalli, S., Tupsamudre, H., Banahatti, V., Lodha, S., 2018. Phishy-a serious game to train enterprise users on phishing awareness. In: *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*. pp. 169–181.
- Cofense, 2022. Keeping up with Phishing. URL: <https://get.cofense.com/CRA-Keeping-Up-With-Phishing.html>.
- CORE, 2021a. Computer science conference rankings descriptions. https://drive.google.com/file/d/1q21YeVEDYyJk9WBpXTgBRH_reCnV12/view.
- CORE, 2021b. CORE conference rankings 2021: Process followed and data considered. <https://drive.google.com/file/d/1bKa40nheaQ3zfuXu3jSpKlw5TnhK9USR/view>.
- Croft, R., Xie, Y., Babar, M.A., 2022. Data preparation for software vulnerability prediction: A systematic literature review. *IEEE Trans. Softw. Eng.*
- Da Silva, F.Q., Santos, A.L., Soares, S., França, A.C.C., Monteiro, C.V., Maciel, F.F., 2011. Six years of systematic literature reviews in software engineering: An updated tertiary study. *Inf. Softw. Technol.* 53 (9), 899–913.
- Das, S., Kim, A., Tingle, Z., Nippert-Eng, C., 2019. All about phishing: Exploring user research through a systematic literature review. *arXiv preprint arXiv:1908.05897*.
- Desolda, G., Ferro, L.S., Marrella, A., Catarci, T., Costabile, M.F., 2021. Human factors in phishing attacks: A systematic literature review. *ACM Comput. Surv.* 54 (8), 1–35.
- Dissanayake, N., Jayatilaka, A., Zahedi, M., Babar, M.A., 2022. Software security patch management-A systematic literature review of challenges, approaches, tools and practices. *Inf. Softw. Technol.* 144, 106771.
- Doshi-Velez, F., Kim, B., 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- Dybå, T., Dingsøyr, T., 2008. Empirical studies of agile software development: A systematic review. *Inf. Softw. Technol.* 50 (9–10), 833–859.
- Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., Beck, H.P., 2003. The role of trust in automation reliance. *Int. J. Hum.-Comput. Stud.* 58 (6), 697–718.
- Egelman, S., Cranor, L.F., Hong, J., 2008. You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. pp. 1065–1074.
- Franz, A., Zimmermann, V., Albrecht, G., Hartwig, K., Reuter, C., Benlian, A., Vogt, J., 2021. (SoK): Still plenty of phish in the sea—A taxonomy of {User-Oriented} phishing interventions and avenues for future research. In: *Seventeenth Symposium on Usable Privacy and Security*. SOUPS 2021, pp. 339–358.
- Garousi, V., Felderer, M., 2017. Experience-based guidelines for effective and efficient data extraction in systematic reviews in software engineering. In: *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering*. pp. 170–179.
- Garousi, V., Felderer, M., Hacıoğlu, T., 2017. Software test maturity assessment and test process improvement: A multivocal literature review. *Inf. Softw. Technol.* 85, 16–42.
- Garousi, V., Felderer, M., Mäntylä, M.V., 2019. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Inf. Softw. Technol.* 106, 101–121.
- Garousi, V., Mäntylä, M.V., 2016. When and what to automate in software testing? A multi-vocal literature review. *Inf. Softw. Technol.* 76, 92–117.
- Greene, K.K., Steves, M., Theofanos, M.F., Kostick, J., et al., 2018. User context: an explanatory variable in phishing susceptibility. In: *Proc. 2018 Workshop Usable Security*.
- Heartfield, R., Loukas, G., 2015. A taxonomy of attacks and a survey of defence mechanisms for semantic social engineering attacks. *ACM Comput. Surv.* 48 (3), 1–39.

- Herzberg, A., Gbara, A., 2004. Trustbar: Protecting (even naive) web users from spoofing and phishing attacks. *Cryptology ePrint Archive, Report 2004/155*. <http://eprint.iacr.org/2004/155>.
- Hu, S., Hsu, C., Zhou, Z., 2021. Security education, training, and awareness programs: Literature review. *J. Comput. Inf. Syst.* 1–13.
- Islam, C., Babar, M.A., Nepal, S., 2019. A multi-vocal review of security orchestration. *ACM Comput. Surv.* 52 (2), 1–45.
- Jampen, D., Gür, G., Sutter, T., Tellenbach, B., 2020. Don't click: towards an effective anti-phishing training. a comparative literature review. *Hum.-Centr. Comput. Inf. Sci.* 10 (1), 1–41.
- Jayatilaka, A., Beu, N., Baetu, I., Zahedi, M., Babar, M.A., Hartley, L., Lewinsmith, W., 2021. Evaluation of security training and awareness programs: Review of current practices and guideline. *arXiv preprint arXiv:2112.06356*.
- Jenkins, A., Kokciyan, N., Vaniea, K.E., 2022. PhishED: Automated contextual feedback for reported Phishing. In: 18th Symposium on Usable Privacy and Security. *Usenix*.
- Jiang, L., Jayatilaka, A., Nasim, M., Grobler, M., Zahedi, M., Babar, M.A., 2021. Systematic literature review on cyber situational awareness visualizations. *arXiv preprint arXiv:2112.10354*.
- Kitchenham, B., 2004. Procedures for Performing Systematic Reviews. Vol. 33. No. 2004. Keele University, Keele, UK, pp. 1–26.
- Kitchenham, B., Charters, S., 2007. Guidelines for Performing Systematic Literature Reviews in Software Engineering. Technical report, EBSE Technical Report EBSE-2007-01, Citeseer.
- Kitchenham, B., Pretorius, R., Budgen, D., Brereton, O.P., Turner, M., Niazi, M., Linkman, S., 2010. Systematic literature reviews in software engineering—a tertiary study. *Inf. Softw. Technol.* 52 (8), 792–805.
- Kumaraguru, P., Rhee, Y., Acquisti, A., Cranor, L.F., Hong, J., Nunge, E., 2007. Protecting people from phishing: the design and evaluation of an embedded training email system. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 905–914.
- Liu, D.-J., Geng, G.-G., Zhang, X.-C., 2022. Multi-scale semantic deep fusion models for phishing website detection. *Expert Syst. Appl.* 209, 118305.
- Magdy, S., Abouelseoud, Y., Mikhail, M., 2022. Efficient spam and phishing emails filtering based on deep learning. *Comput. Netw.* 206, 108826.
- Mahood, Q., Van Eerd, D., Irvin, E., 2014. Searching for grey literature for systematic reviews: challenges and benefits. *Res. Synth. Methods* 5 (3), 221–234.
- Metaxas, P.T., 2009. Web spam, social propaganda and the evolution of search engine rankings. In: International Conference on Web Information Systems and Technologies. Springer, pp. 170–182.
- Petelka, J., Zou, Y., Schaub, F., 2019. Put your warning where your link is: Improving and evaluating email phishing warnings. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 1–15.
- Reuter, C., Iacono, L.L., Benlian, A., 2022. A quarter century of usable security and privacy research: transparency, tailorability, and the road ahead. *Behav. Inf. Technol.* 1–14.
- Sabir, B., Ullah, F., Babar, M.A., Gaire, R., 2021. Machine learning for detecting data exfiltration: a review. *ACM Comput. Surv.* 54 (3), 1–47.
- Salamah, F.B., Palomino, M.A., Papadaki, M., Furnell, S., 2022. The importance of the job role in social media cybersecurity training. In: 2022 IEEE European Symposium on Security and Privacy Workshops. EuroS&PW, IEEE, pp. 454–462.
- Sbaraini, A., Carter, S.M., Evans, R.W., Blinkhorn, A., 2011. How to do a grounded theory study: a worked example of a study of dental practices. *BMC Med. Res. Methodol.* 11 (1), 1–10.
- Schechter, S.E., Dhamija, R., Ozment, A., Fischer, I., 2007. The emperor's new security indicators. In: 2007 IEEE Symposium on Security and Privacy. SP'07, IEEE, pp. 51–65.
- Schuetz, S.W., Steelman, Z.R., Syler, R.A., 2022. It's not just about accuracy: An investigation of the human factors in users' reliance on anti-phishing tools. *Decis. Support Syst.* 113846.
- Shahin, M., Babar, M.A., Chauhan, M.A., 2020. Architectural design space for modelling and simulation as a service: a review. *J. Syst. Softw.* 170, 110752.
- Shahin, M., Babar, M.A., Zhu, L., 2017. Continuous integration, delivery and deployment: a systematic review on approaches, tools, challenges and practices. *IEEE Access* 5, 3909–3943.
- Sharma, G., 2021. Fight the phish—see how Microsoft Learn can help. <https://techcommunity.microsoft.com/t5/microsoft-learn-blog/fight-the-phish-see-how-microsoft-learn-can-help/ba-p/2824122>.
- Sheng, S., Magnien, B., Kumaraguru, P., Acquisti, A., Cranor, L.F., Hong, J., Nunge, E., 2007. Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. In: Proceedings of the 3rd Symposium on Usable Privacy and Security. pp. 88–99.
- Sjouwerman, S., 2021. Context is the key to phishing success. URL: <https://blog.knowbe4.com/context-is-the-key-to-phishing-success>.
- Soneji, A., Kokulu, F.B., Rubio-Medrano, C., Bao, T., Wang, R., Shoshitaishvili, Y., Doupe, A., 2022. “Flawed, but like democracy we don't have a better system”: The Experts' Insights on the Peer Review Process of Evaluating Security Papers. In: 2022 IEEE Symposium on Security and Privacy. SP, IEEE, pp. 1845–1862.
- Souza, E., Moreira, A., Goulão, M., 2019. Deriving architectural models from requirements specifications: A systematic mapping study. *Inf. Softw. Technol.* 109, 26–39.
- Szathmari, G., 2019. Owning your legal practice's cybersecurity – MSPs and you. URL: <https://get.cofense.com/CRA-Keeping-Up-With-Phishing.html>.
- Teraguchi, N.C.R.L.Y., Mitchell, J.C., 2004. Client-Side Defense Against Web-Based Identity Theft. Computer Science Department, Stanford University, Available: <http://crypto.stanford.edu/SpoofGuard/webspoof.pdf>. Citeseer.
- Terranova Security, 2022. How to avoid phishing simulations false positives? URL: <https://terranovasecurity.com/phishing-simulations-false-positives/>.
- Tom, E., Aurum, A., Vidgen, R., 2013. An exploration of technical debt. *J. Syst. Softw.* 86 (6), 1498–1516.
- Vilone, G., Longo, L., 2020. Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*.
- Zahedi, M., Shahin, M., Babar, M.A., 2016. A systematic review of knowledge sharing challenges and practices in global software development. *Int. J. Inf. Manage.* 36 (6), 995–1019.

Orvila Sarker is a Ph.D. candidate enrolled in the School of Computer Science at The University of Adelaide, Australia. Additionally, she holds an Assistant Professor position in the Department of Information and Communication Technology at Comilla University in Bangladesh. Her research focuses on the development of enhanced approaches to incorporate individual end user requirements into the design of anti-phishing interventions for the purposes of phishing education, training, and awareness. This entails the identification of human-centric and socio-technical aspects associated with existing anti-phishing interventions, with the goal of educating, training, and increasing individuals' awareness regarding the prevention of phishing attacks. The anticipated outcome of her doctoral research is the creation of tools or guidelines to facilitate a more comprehensive understanding and mitigation of human-centric challenges during the design, implementation, and evaluation stages of present anti-phishing technologies.

Asangi Jayatilaka received the Ph.D. degree from the School of Computer Science, The University of Adelaide (UoA). She is currently a research fellow at the Centre for Research on Software Technologies (CREST), UoA. Prior to joining CREST, she worked as a Lecturer in software engineering and computer science at the School of Computer Science, UoA. She is passionate about research on human aspects in computing. This includes studying the effects of different human aspects on technology development, whether and/or to what extent these are accounted for and how we can best use these to build better tools and technologies that are both usable and effective. She has extensive experience in both qualitative and quantitative research methods. Her work has led to designing, implementing, and evaluating technologies and tools in various domains, including cyber security, digital health, and pervasive computing.

Sherif Haggag is an Assistant Professor in School of Computer Science at the University of Adelaide. He has worked at various universities, such as Deakin University, where he achieved his PhD. His research areas of interest include Human-Centred Software Engineering, Cybersecurity, Understanding Human-Centric issues and designing apps with adaptive user interfaces, Human factors and social engineering in Cybersecurity and the persistence of the Privacy Paradox and cybersecurity behavior. Sherif strongly believes that software engineering is designed to solve human problems and support humans in different aspects, such as health, education, transport, manufacturing, etc. However, current software engineering techniques do not pay attention to other humans/end users who use the same system.

Chelsea Liu is an Associate Professor at the Adelaide Business School, The University of Adelaide. She has previously worked as a Visiting Assistant Professor in Finance at the University of Washington. Chelsea holds a Master of Computer and Information Technology from the University of Pennsylvania. She also has a Ph.D. in Business from The University of Adelaide and is qualified as a barrister and solicitor in Australia. Chelsea's interdisciplinary research focuses on the grand challenges facing our society including cybersecurity and climate change. Chelsea's research has been published in leading academic journals and presented globally including at Cornell, Northwestern, Stanford, Duke Universities, and University of Washington, among others. Chelsea currently serves as a Vice President of Australia's Financial Research Network (FIRN).

M. Ali Babar is currently a Professor with the School of Computer Science, The University of Adelaide. He is an Honorary Visiting Professor with the Software Institute, Nanjing University, China. He is also the Director of the Cyber Security Adelaide (CSA), which incorporates a node of recently approved the Cyber Security Cooperative Research Centre (CSCRC), whose estimated budget is around AU\$140 million over seven years with AU\$50 million provided by the Australia Government. In the area of software engineering education, he led the university's effort to redevelop a Bachelor of Engineering (software) degree that has been accredited by the Australian Computer Society and the Engineers Australia (ACS/EA). Prior to joining The University of Adelaide, he spent almost seven years in Europe (Ireland, Denmark, and U.K.) as a Senior Researcher and an Academician. Before returning to Australia, he was a Reader of software engineering with Lancaster University. He has established

an Interdisciplinary Research Centre, Centre for Research on Engineering Software Technologies (CREST), where he leads the research and research training of more than 30 (20 Ph.D. students) members. Apart from his work having industrial relevance as evidenced by several research and development projects and setting up several collaborations in Australia and Europe with industry and government agencies, his

publications have been highly cited within the discipline of software engineering as evidenced by his H-index is 52 with 11045 citations as per Google Scholar (December 2021). He leads the theme on Platform and Architecture for Cyber Security as a Service with CSCRC. He has authored/co-authored more than 220 peer reviewed publications through premier software technology journals and conferences.