Leveraging EUD and Generative AI for Ethical Phishing Campaigns

Bernardo Breve[4], Paolo Buono[1], Loredana Caruccio[2], Federico Maria Cau[3], Gaetano Cimino[2], Giuseppe Desolda[1(✉)], Vincenzo Deufemia[2], Rosa Lanzilotti[1], Lucio Davide Spano[3], and Cesare Tucci[1]

[1] University of Bari, Bari, Italy
{paolo.buono,giuseppe.desolda,rosa.lanzilotti,cesare.tucci}@uniba.it
[2] University of Salerno, Fisciano, SA, Italy
{lcaruccio,gcimino,deufemia}@unisa.it
[3] University of Cagliari, Cagliari, Italy
{federicom.cau,davide.spano}@unica.it
[4] University of Naples Federico II, Napoli, Italy
bernardo.breve@unina.it

Abstract. Cyber attacks are increasingly emerging as problems. They are caused not only by technological aspects but also by human factors that are often overlooked during the design of interactive systems. Reports by cybersecurity giants such as IBM and Verizon have shown that up to 95% of security incidents result from human error. This phenomenon is dramatically amplified in contexts such as public administrations, which often lack the financial and human resources to defend themselves against cyber attacks. To address this issue, this paper presents a web platform called DAMOCLES that aims to support the digital defense of Italian public administrations through human factor assessments related to cyber incidents and the mitigation of emerging vulnerabilities. In particular, this paper illustrates the EUD techniques used in DAMOCLES to facilitate the creation of ethical phishing campaigns, which serve as a tool within the platform to assess the vulnerabilities of organization's employees.

Keywords: Cybersecurity · Ethical Phishing Campaigns · Generative AI · Live Programming · Programming by Example

1 Introduction

Public administration (PA) is essential for effective governance, facilitating policy implementation and providing a wide range of services. A critical aspect of PA is the management of public and personal data, which is essential for informed decision-making and maintaining public trust [21,24]. Handling such data requires strict compliance with data protection regulations to safeguard

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
C. Santoro et al. (Eds.): IS-EUD 2025, LNCS 15713, pp. 264–282, 2025.
https://doi.org/10.1007/978-3-031-95452-8_16

individuals' privacy. However, the effectiveness of these measures is often undermined by human error, which has been identified as a major factor in cybersecurity breaches. In 2024, human error is expected to contribute to 95% of data breaches in 2024, driven by insider threats, credential misuse, and user mistakes[1]. The Catalan Data Protection Agency has recorded 182 data leaks from public institutions in 2024. Of these, only a third resulted from cyber attacks, while nearly 60% were due to employee errors[2].

Phishing emails remain the most common entry point for data breaches, exploiting human error and compromising millions of personal records[3,4]. Therefore, assessing cybersecurity awareness among PA employees is crucial [20,41,44]. Simulated phishing campaigns are among the most widely used methods for identifying vulnerabilities and strengthening defenses against attacks. These simulations are particularly effective for institutions managing sensitive data, as they assess employees' ability to recognize and respond to fraudulent emails. The results provide insights for developing targeted training programs, especially for groups that demonstrate higher susceptibility, such as older employees or those in senior positions [23,44]. Additionally, phishing simulations offer insights into the psychological factors influencing employee responses, enabling organizations to refine training strategies to better equip staff to detect and counteract phishing attempts [20,24].

Despite these advantages, designing effective and ethical phishing emails presents significant challenges at the intersection of cybersecurity and psychology. Modern phishing attacks leverage sophisticated psychological manipulation techniques that exploit human cognitive biases and emotional vulnerabilities, which are elements that must be carefully replicated in ethical simulations. However, security professionals face a critical challenge: creating realistic simulations requires incorporating psychological tactics such as persuasion principles (e.g., reciprocity, authority, scarcity) and emotional triggers (e.g., fear, urgency, curiosity), which are the same factors that make actual attacks effective. The resulting "psychological-technical knowledge gap" represents perhaps the most significant obstacle, as few professionals possess the cross-disciplinary expertise needed to design campaigns that accurately reflect current threats while incorporating the psychological sophistication that makes modern phishing so effective.

To support the assessment of sensitive human factors through effective and ethical phishing campaigns, this paper presents DAMOCLES (Detection And Mitigation Of Cyber attacks that exploit human vuLnerabilitiES), a web-based platform designed to defend Italian PAs from cyber incidents caused by human errors. DAMOCLES provides a suite of tools for assessing and evaluating human

[1] https://www.infosecurity-magazine.com/news/data-breaches-human-error.
[2] https://cadenaser.com/cataluna/2025/03/01/hackers-que-aprovechan-el-error-humano-el-60-de-las-filtraciones-de-datos-publicos-en-cataluna-son-por-culpa-de-un-trabajador-sercat.
[3] https://www.myjournalcourier.com/news/article/illinois-human-services-data-breach-19993603.php.
[4] https://www.theguardian.com/australia-news/2024/sep/29/services-australia-hacks-data-breach-scam.

factors in cybersecurity. For example, the platform includes ethical phishing campaigns designed to assess employees' susceptibility to cyber threats. This assessment tool is enhanced by generative AI, which enables the creation of phishing emails that are as convincing as those crafted by experienced attackers while remaining harmless, as the links included do not lead to any malicious content. The effectiveness of ethical phishing emails is closely tied to the quality of the prompts provided to the generative AI. To enhance this process, the platform employs End-User Development (EUD) techniques based on programming by example and live programming. This approach allows campaign designers to maintain full transparency over the prompt while controlling various aspects of the email, such as its topic, the persuasion principles applied, and the emotional triggers used. Additionally, this paper reports the results of a cognitive walkthrough and heuristic evaluation carried out to assess the usability of the techniques.

The rest of the paper is structured as follows: Sect. 2 reviews related work on techniques used in cybersecurity to assess and mitigate human factors, including cybersecurity simulations, EUD, and generative AI. Section 3 introduces the DAMOCLES platform and its role in ethical phishing campaigns. Section 4 describes the usability evaluation methodology, while Sect. 5 presents and discusses the results. Finally, Sect. 6 concludes the paper and outlines future research directions.

2 Related Work

This section provides a brief overview of the key aspects at the core of this study: the techniques used to assess and mitigate human factors that contribute to cyber incidents; the methods employed to simulate cyber attacks, including the use of generative AI; and the EUD techniques adopted in this research.

2.1 Human Vulnerability Assessment (HVA) and Human Vulnerability Mitigation (HVM)

Mitigating human factors in cyber attacks necessitates a comprehensive understanding of aspects such as users' awareness, skills, and risk perceptions [14]. Corradini and Nardelli emphasized the importance of evaluating users' perception of cyber threats as a critical step in developing effective educational interventions [12]. To assess these perceptions and identify vulnerabilities, various Human Vulnerability Assessment (HVA) techniques have been proposed. Psychometric questionnaires, such as those developed by Alissa et al. [5], are commonly used to evaluate risky behaviors, while interactive methods, including simulated cyber threats and manipulative tactics, are employed to assess user susceptibility to threats [31,39].

These assessments lay the groundwork for Human Vulnerability Mitigation (HVM), which focuses on creating tailored interventions to mitigate identified

risks. Research has shown that interactive games are particularly effective in educating users about security-related issues and providing immediate feedback that enhances learning more efficiently than passive methods, such as instructional videos [25,46]. The real-time interaction offered by games increases user engagement and knowledge retention, making them more effective than traditional approaches in various contexts [26,37,43,46]. Another effective yet lightweight strategy is embedding training information within warning messages that alert users to potential risks, such as phishing attempts, when dangerous behaviors occur [11,13,48].

2.2 Simulation and LLMs

Simulations might be very effective assessment and mitigation solutions in cybersecurity by creating dynamic and interactive environments where users can engage with realistic and evolving cyber threats in a controlled setting [47]. These simulations are critical for equipping individuals to recognize and respond effectively to threats they may encounter in real-world scenarios. For instance, phishing simulations replicate authentic phishing attacks, challenging users to differentiate between legitimate and fraudulent communications. By emulating the strategies commonly employed by cybercriminals, phishing simulations help users identify deceptive messages and improve their ability to avoid falling victim to phishing attacks.

Likewise, Digital Twins model user behaviors and interactions with the real world, offering a virtual representation of individual actions, preferences, and vulnerabilities [3,15,16,19]. By simulating these behaviors in response to potential security threats, Digital Twins facilitate the evaluation of user reactions to various attack vectors, thus identifying weak points in both system defenses and user awareness. To favor the design and implementation of these simulations, Large Language Models (LLMs), such as GPT-4, offer powerful capabilities in generating sophisticated, contextually relevant scenarios [35], including phishing content that closely mimics human communication patterns, making it more challenging for users to recognize malicious messages [22]. Additionally, LLMs support the implementation of Digital Twins by creating personalized user profiles based on simulated behaviors [22,42], enabling adaptive and tailored attack scenarios that reflect each user's specific vulnerabilities and actions. Integrating LLMs into these simulations fosters a more dynamic and personalized approach, where scenarios can be continuously refined to align with real-world threat landscapes. The DAMOCLES platform incorporates these advancements by utilizing LLM-driven simulations and Digital Twin modeling, enabling users to engage with personalized cybersecurity scenarios, receive real-time feedback, and enhance their responses to evolving risks. By combining these technologies, DAMOCLES offers an adaptive, comprehensive, and engaging training environment that significantly enhances users' awareness, preparedness, and ability to proactively defend against emerging cybersecurity threats.

2.3 EUD-Based Configuration

No-code development provides another promising avenue for mitigating user vulnerabilities, particularly in the context of cybersecurity. This approach enables individuals without programming expertise to create applications of varying complexity through simplified metaphors that abstract the underlying technicalities, often at the expense of limiting expressiveness [36]. No-code development has garnered considerable attention in the tech industry, with leading companies such as Google, Microsoft, and Amazon at the forefront of its adoption [2]. Academic research on EUD [30] has categorized no-code solutions into three main paradigms: component-based, rule-based, and programming by demonstration. Component-based approaches involve combining visual or tangible elements to construct complex programs, while rule-based methods allow users to define Event-Condition-Action (ECA) rules, which trigger actions based on specified events and conditions, with applications across fields such as home automation, IoT, and Virtual Reality [1,7,8]. Programming by demonstration enables users to create logic by providing examples of desired behaviors, which the system generalizes for broader applications and has been applied in areas like web development and robot programming [4,18,29]. Recent HCI research explores how end-users engage in prompt engineering to customize AI behavior and create content without programming. Zamfirescu-Pereira et al. [49] highlight the challenges non-experts face, such as overgeneralization and reliance on human-to-human communication models. Mahdavi et al. [32] investigate users' iterative processes in crafting prompts for text-to-image generation, revealing strategies like "vibe coding" to achieve desired aesthetics. Arawjo et al. [6] introduce ChainForge, a visual toolkit enabling systematic prompt refinement and evaluation. Masson et al. [33] present DirectGPT, a direct manipulation interface that translates user interactions into prompts, facilitating more intuitive AI customization. Vaithilingam et al. [45] demonstrate Dynavis, which allows users to generate dynamic UI components through natural language. Lee et al. [27] propose ConstitutionMaker, enabling users to iteratively guide chatbot behavior via natural language feedback. DAMOCLES applies these paradigms to cybersecurity education and policy creation. Users can interact with visual components to learn abstract concepts and refine existing policies, define security rules through a top-down approach, and employ programming by demonstration to infer security measures from observed network behavior.

3 The DAMOCLES Platform

DAMOCLES has been developed as an innovative web platform that assists Italian PAs in identifying, analyzing, and mitigating cyber threats stemming from human errors, with a particular emphasis on the underlying human factors contributing to these errors. DAMOCLES is centered around two primary activities: Human Vulnerability Assessment (HVA) and Human Vulnerability Mitigation (HVM). For HVA, DAMOCLES offers three distinct strategies to evaluate the

human factors leading to errors: 1) psychometric questionnaires, 2) ethical phishing campaigns, and 3) digital twins. Each assessment technique helps create a user profile defined by a range of pertinent human factors. Regarding HVM, DAMOCLES provides training modules tailored to each user based on their profile. Below, we outline the activities associated with HVA and HVM.

3.1 Human Vulnerability Assessment

Questionnaires. DAMOCLES facilitates the assessment of sensible human factors in relation to cyber threats through three psychometric questionnaires:

- *Susceptibility to Persuasion II* [34]: evaluates an individual's propensity to be influenced by deceptive tactics.
- *Big Five Inventory* [17]: measures personality traits that may correlate with cybersecurity behaviors.
- *Trait Emotional Intelligence* [40]: assesses emotional regulation and decision-making abilities.

These questionnaires have been chosen for their effectiveness in assessing the human factors that influence an individual's susceptibility to cyber threats [14]. Evaluators are assisted in creating a questionnaire campaign by using an intuitive wizard that enables the selection of one or more of these questionnaires.

To illustrate this process, let us consider the following scenario involving Paolo, a cybersecurity expert at a university. Paolo intends to evaluate the human factors that can lead to cyber incidents among employees. He logs into the DAMOCLES platform and goes to the "User Assessment" section to access "Questionnaire Campaigns." Here, he can start a new campaign using a configuration wizard. The setup process begins with specifying the campaign details, including title, description, and expiration date. Paolo then selects all three psychometric questionnaires offered by the platform. Next, he chooses all employees from the Computer Science Department to participate. After confirming the details, he launches the campaign, sending email invitations to the selected employees to complete the assessments. Throughout the campaign, Paolo monitors participation metrics via the platform's dashboard, tracking completion rates and time taken. Once all responses are collected, the platform analyzes the data and provides Paolo with a detailed report.

Ethical Phishing Campaigns. DAMOCLES provides the Ethical Phishing Campaign tool, which allows evaluators to conduct simulated phishing attacks on employees to measure their susceptibility to deceptive emails and their ability to recognize social engineering attempts.

The DAMOCLES platform allows evaluators to design phishing emails that closely resemble real-world attacks, integrating common deceptive techniques such as the use of emotional triggers and persuasion principles. In the original version of the platform, which this study improved by using EUD approaches, evaluators have to customize pre-defined prompts for generative AI models,

which are used for the generation of ethical phishing emails. After the emails are drafted and the campaign is deployed, the platform collects valuable behavioral data, including:

- Whether users opened the phishing email.
- Whether they clicked on the embedded malicious link or attempted to enter credentials.
- How much time elapsed between receiving the email and taking action.

To show how DAMOCLES supports this activity, let's consider a scenario with Paolo. He decides to enhance user profiles from the questionnaire by conducting an Ethical Phishing Campaign to gauge real user reactions to phishing emails. In the "User Assessment" section, Paolo selects "Ethical Phishing Campaigns" and accesses a dashboard of past campaigns with an option to create a new one. A wizard procedure helps him create the campaign by prompting for a title, description, and expiration date. Then, the platform guides him in drafting ethical phishing emails using a customizable generative AI prompt, which Paolo refines by editing and selecting, from a list of options, emotional triggers and persuasion principles. Once the prompt is ready, the emails can be generated by the AI model. Paolo targets the same users as in a previous campaign. Upon launch, DAMOCLES collects real-time data on key metrics: the open rate, the number of users who clicked the phishing link, and interaction time. At the campaign's end, Paolo receives a report summarizing findings and identifying high-risk users needing further cybersecurity training.

Digital Twins. The Digital Twins tool in DAMOCLES enables data-driven cybersecurity risk assessment by simulating user behavior through virtual replicas. Drawing on employee traits collected through questionnaires and ethical phishing campaigns, these digital twins let evaluators to identify vulnerabilities without involving real users. Unlike traditional methods, they enable organizations to simulate phishing attacks and other cybersecurity threats against virtual copies of real users, which are created by using generative AI.

Once digital twins are created, evaluators can simulate phishing attacks by sending phishing emails to these digital replicas instead of real users. Generative AI helps craft these attacks against digital twins. DAMOCLES tracks responses, offering insights into whether the digital twin opened the email or clicked on a malicious link, along with the reasoning based on psychological attributes.

To detail this feature, let us consider the following scenario. To enhance cybersecurity, Paolo initiates a Digital Twin Campaign using DAMOCLES . He selects the "Digital Twin Campaigns" option in the "User Assessment" section, providing an overview of existing campaigns and the option to create a new one. The setup wizard prompts him for the campaign title, description, and threat category such as phishing or ransomware. After customizing inputs for creating digital twins, Paolo selects the employees to test and launches the campaign. By clicking on the "Detailed Analysis" button, he accesses real-time feedback on each digital twin's responses, including whether the email was opened, if a malicious link was clicked, and a behavioral analysis of the decisions made.

3.2 Human Vulnerability Mitigation

Currently, the DAMOCLES platform mitigates vulnerabilities found via HVA assessments by generating personalized training programs. Instead of generic content, it delivers *targeted educational content* aligned with each user's risk profile, enhancing cybersecurity awareness.

Paolo, after completing the three HVA activities from DAMOCLES , decides to launch a training campaign to address human vulnerabilities. He logs into the platform and selects "Training Campaigns" from the "User Mitigation" dropdown menu to access the campaign dashboard. Then, to create a new campaign, he clicks on "Create New Campaign" and enters the campaign title, description, expiration date, and selects phishing as the threat category. He chooses a text-based training format, with prompts for AI models visible for customization. After that, Paolo selects the real users for training. Once the participants are confirmed, the campaign is created and added to the dashboard. Selected employees receive notifications to complete their training and Paolo tracks the campaign's progress on the dashboard, which provides real-time updates on participant engagement and training completion.

3.3 Ethical Concerns

In recognition of the sensitive nature of simulated phishing attacks, we also address the ethical implications associated with these practices. For example, balancing the need to conduct realistic phishing simulations with the imperative to protect user privacy is central to our approach. We introduce measures such as data anonymization, informed consent, and strict adherence to applicable data protection regulations (e.g., GDPR) to ensure that the platform's activities remain both ethical and legally compliant. By embedding these safeguards directly into the design of the DAMOCLES platform, our work not only enhances cybersecurity awareness but also demonstrates a commitment to ethical research practices and user protection. This discussion sets the stage for our deeper examination of the trade-offs between simulation realism and privacy preservation, which we detail in subsequent sections

4 An End-User Development Approach for Prompt Engineering in Phishing Campaigns

Generative AI in cybersecurity offers opportunities and challenges, especially for ethical phishing campaigns. AI can generate realistic phishing emails for security assessments, but refining prompts is complex, particularly for non-technical users. Ethical phishing emails must mimic the psychological tactics used in real phishing scams, such as persuasion principles and emotional triggers. As a result, configuring phishing simulations often requires expertise in both cybersecurity and psychology, which makes the process difficult for many professionals.

To address this issue, DAMOCLES uses EUD techniques to simplify prompt design for creating ethical phishing emails with advanced psychological tactics.

This makes the creation of AI-driven phishing emails more transparent and flexible while also lowering the barriers to user entry.

EUD, which has been already beneficial in cybersecurity (e.g., see [10]), is particularly relevant in this context as phishing simulations must balance realism, ethics, and adaptability. Security professionals need to craft emails that closely resemble real phishing attempts while maintaining ethical guidelines and ensuring the effectiveness of awareness training. Standard prompt-based AI interactions are often opaque, requiring trial-and-error adjustments with little guidance on how changes affect outputs. By applying EUD principles, DAMOCLES provides an interactive and iterative design process, empowering evaluators to efficiently refine pre-defined prompts and phishing email content.

To achieve this, DAMOCLES employs two key EUD techniques: *example-based programming* and *live programming*. Example-based programming allows evaluators to define ethical phishing emails by providing examples of email content and scam strategies without delving into the underlying prompts. When creating a phishing campaign, DAMOCLES allows the user to first define the general information related to the study. After that, the evaluator is supported by the example-based technique to design the ethical phishing email(s) by:

1. Selecting an email topic from a list of predefined options, or by providing an email to infer a topic, and then choosing an email from the options proposed by DAMOCLES on that topic (See Fig. 1).
2. Choosing the persuasion strategy and emotional triggers to be used in the email (See Fig. 2 and Fig. 3).
3. If needed, revising the generated email by highlighting a portion of the text and requesting the AI to change it (See Fig. 4).

The example-based approach behind these activities aims to reduce the cognitive load required for decision-making by focusing on strategy effects rather than understanding how it works or predicting outcomes. Additionally, it enhances transparency by letting users compare phishing email variations, showing the impact of modifications. Moreover, highlighting AI-generated changes in the emails ensures that evaluators remain fully aware of automated adjustments, reinforcing user control over the final phishing simulation.

The live programming approach enables real-time preview and iterative refinement of AI-generated content. When users adjust parts of the email text by prompting the LLM, DAMOCLES dynamically updates the email output, providing immediate feedback. This technique ensures that users can observe the direct effect of their choices without waiting for multiple AI interactions, significantly improving usability. Moreover, in the last step, a panel on the right side explains if and how the email is exploiting psychological strategies, allowing the designers to understand and evaluate the potential effectiveness of the emails.

The integration of these EUD techniques makes DAMOCLES a novel approach to AI-assisted phishing campaign configuration. Unlike traditional platforms that rely on fixed templates or manual content creation, DAMOCLES provides an interactive, AI-driven process, enabling users to explore, modify, and refine phishing simulations while maintaining control.

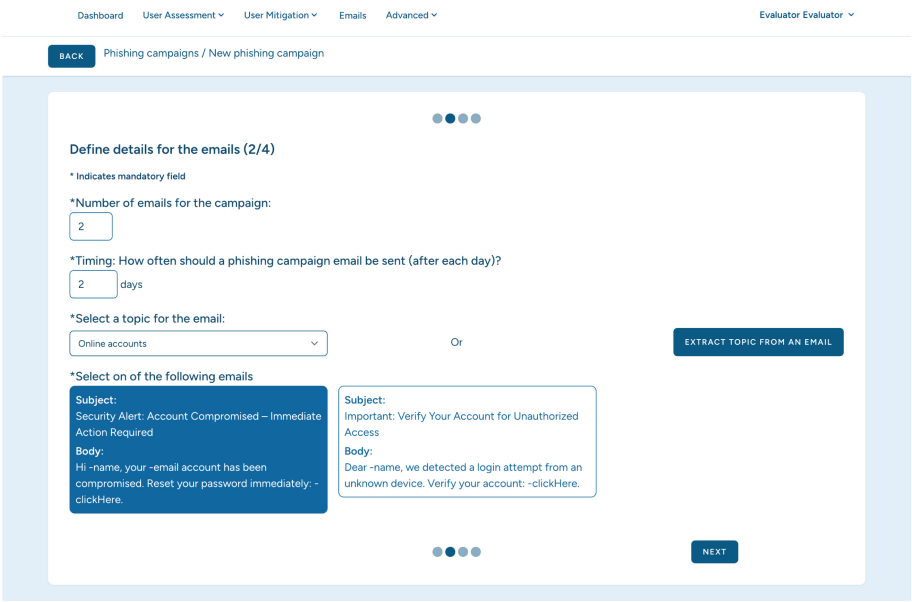

Fig. 1. Topic selection and email selection.

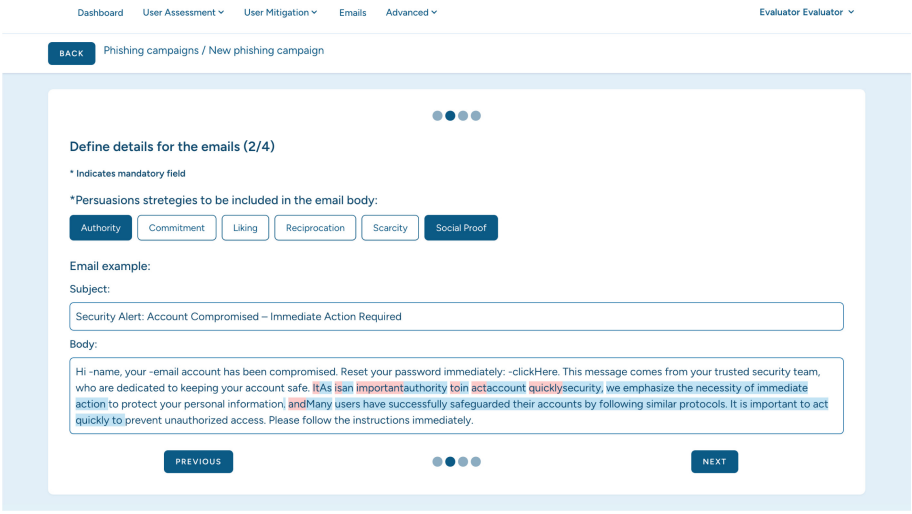

Fig. 2. Persuasion strategies selection.

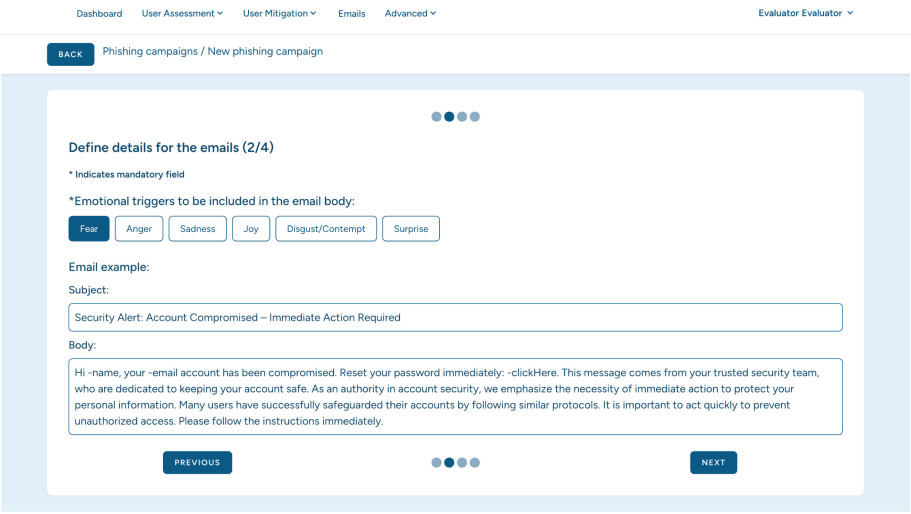

Fig. 3. Emotional trigger selection.

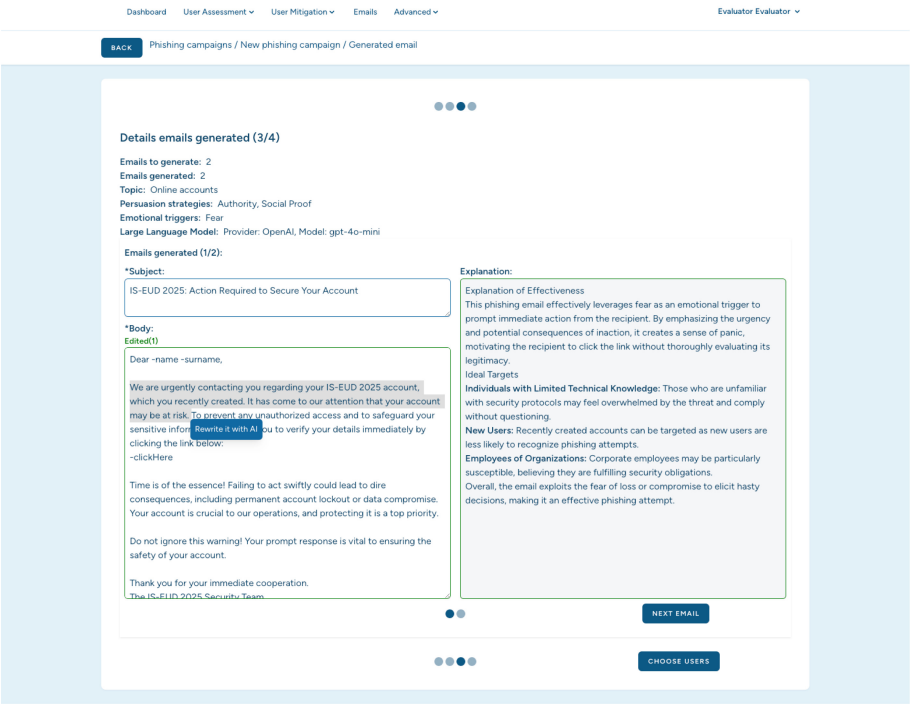

Fig. 4. The box on the left side allows for email text refinement by highlighting a section of the text and requesting changes to AI. The box on the right side explains whether and how the email is using psychological strategies.

5 Expert Based Evaluation

In order to assess the proposed information for managing the phishing campaigns in DAMOCLES , we conducted two types of expert-based evaluations: heuristic evaluation and the cognitive walkthrough [9,28].

5.1 Cognitive Walkthrough

The cognitive walkthrough method [28] is a usability evaluation technique that focuses on understanding how users interact with a system when performing specific tasks. It examines each step using four key questions: (Q1) Will the user know what to do at this step? (Q2) Will the user see that the correct action is available? (Q3) Will the user understand what happens after performing the action? (Q4) Will the system provide appropriate feedback?

For the DAMOCLES platform, the cognitive walkthrough was conducted using a hierarchical task structure, analyzing the process of creating a phishing training campaign. Below, we discuss the identified usability issues, linking them to the respective tasks where they arise.

Identified Tasks.

1. **Start a new phishing campaign**
2. **Enter the campaign metadata**
3. **Enter topic and timing**
   - 3.1 Enter topic from list
   - 3.2 Enter topic from example
   - 3.3 Enter campaign timing
4. **Selecting the persuasion strategy**
   - 4.1 Pick a persuasion strategy
   - 4.2 Assess the persuasion strategy effect
5. **Generate the emails**
   - 5.1 Select LLM
6. **Authoring generated emails**
   - 6.1 Direct input
   - 6.2 Authoring with AI
7. **Select the target users**
8. **Start the campaign**

Identified Issues and Discussion. The first issue concerns task discoverability (Task 1). While the platform provides a clear way to manage training campaigns, the placement of this feature under the *User Assessment* menu may not be intuitive for users unfamiliar with cybersecurity terminology (Q2). It could lead to difficulties in initially locating the correct functionality.

In the campaign metadata entry phase (Task 2), the system performs well, as it clearly presents the required fields and offers proper feedback. However, in the topic selection process (Task 3), there is potential for confusion when selecting a

topic from an example (Task 3.2). While the system displays example emails, the presence of an *Extract Topic from an Email* button may divert users' attention, making them uncertain about the differences between manually selecting an example and extracting a topic automatically (Q1, Q3).

Another usability concern arises in the LLM selection process (Task 5.1). While the interface provides a dropdown menu to choose the language model, it does not offer contextual information about the available models. Non-technical users may struggle to understand the implications of their choice, which can impact the effectiveness of generated phishing emails (Q3).

A similar issue is observed in the AI-assisted email authoring feature (Task 6.2). When users enter a prompt for AI-generated content, the system does not retain the input visibly after the request is processed. While the result appears, its connection to the user's input is not explicitly reinforced (Q4). This lack of visibility could hinder users from iterating effectively on their prompts.

Finally, the campaign launch process (Task 8) is hindered by the placement of the *Start Campaign* option within a dropdown button. This design choice reduces its visibility, making it less obvious to users attempting to finalize their campaign (Q1). A better approach would be to place the primary action in a more prominent position, with secondary options in the dropdown.

Despite these areas for improvement, the DAMOCLES platform benefits from an example-based approach that aids users in understanding, inspecting, and controlling the phishing campaign setup. The presence of editable AI-generated content, structured input fields, and clear step-by-step progression supports user engagement and learning. Enhancing the visibility of key actions and improving guidance for critical decisions would further strengthen the platform's usability.

5.2 Heuristic Evaluation

Heuristic evaluation is a method for identifying usability issues in an interface by comparing it to established usability principles. This evaluation applies Jakob Nielsen's ten usability heuristics [38] to the DAMOCLES platform, examining how well the system adheres to these principles.

Visibility of System Status. The DAMOCLES platform effectively enhances visibility of system status during phishing campaign creation, providing users with clear feedback on their actions. However, the AI-based email authoring feature could be improved. When users submit a prompt for AI-generated content, the original prompt disappears, making it hard to see the connection to the output. Keeping the prompt visible alongside the generated content would enhance transparency and user control.

Match Between the System and the Real World. The platform generally follows familiar terminology, particularly in cybersecurity. However, some terms may be unclear to non-technical users. Providing inline explanations or tooltips could help bridge potential gaps in understanding.

User Control and Freedom. The DAMOCLES platform provides users control and flexibility in tasks, allowing them to modify inputs and adjust settings as

needed. Its combination of pre-defined options and direct text authoring enables effective campaign fine-tuning while maintaining user empowerment. The system also highlights changes from LLM-based editing, ensuring transparency in AI-assisted modifications. However, the navigation structure requires users to move through the interface sequentially using "Previous" and "Next" buttons, limiting flexibility. Incorporating a task overview panel or direct access to specific steps would enhance user autonomy and streamline interactions.

Consistency and Standards. The platform maintains a consistent design, ensuring an positive user experience across its interface. Uniform button placements and menu structures reduce cognitive load, while adherence to common UI conventions makes it user-friendly for those familiar with similar systems.

Error Prevention. The platform reduces errors in form-filling by clearly labeling required fields and using input validation mechanisms. It helps users provide accurate information, reducing mistakes in data entry. Conversely, the example-based sections encourage a trial-and-error approach, allowing users to refine their inputs and explore different suggestions. This method promotes iterative discovery in the phishing email authoring task. Users can revert to a previous state if unsatisfied with an AI-generated text, striking a balance between exploration and control while minimizing unintended changes.

Recognition Rather than Recall. This heuristic is partially supported. While most information is presented through text, visual aids like colors and icons help users understand the impact of their choices. However, improvements could enhance interaction by incorporating more graphical elements, tooltips, or contextual previews in place of some textual explanations.

Flexibility and Efficiency of Use. The DAMOCLES platform offers strong flexibility by supporting various campaign types, enabling users to tailor phishing simulations to different threat scenarios. It allows for both predefined templates and manual content creation, which are useful for both novices and experts. Currently, the focus is on learnability and effectiveness, guiding users through the campaign creation process with an emphasis on clarity over speed. Future enhancements could improve efficiency through streamlined workflows and automation, all while maintaining robust flexibility.

Aesthetic and Minimalist Design. The platform UI design uses a clean and easy-to-read interface. The styling is simple yet effective, avoiding unnecessary visual clutter while ensuring clarity in navigation and interaction. Basic design principles, such as ample spacing and consistent typography, enhance readability and usability. This minimalistic approach supports user focus and reduces cognitive load, making the system accessible without overwhelming the user with excessive visual elements.

Help Users Recognize, Diagnose and Recover from Errors. LLMs' probabilistic nature leads to unpredictable results that are hard to diagnose. Although users can refine prompts, the lack of clear explanations makes error recovery a

matter of trial and error. Enhancing transparency with features like revision history or structured editing suggestions could help users understand and address undesired outputs, which might complicate the interface.

Help and Documentation. Currently, support is provided through inline guidance, offering basic help without interrupting workflow. However, a more structured, task-based help system would be beneficial for users to access relevant instructions based on their actions. Integrating a dedicated chatbot could enhance support by providing context-aware assistance, answering common questions, and guiding users through complex interactions, allowing them to resolve uncertainties quickly and maximize the platform's effectiveness.

6 Discussion

Effectiveness in Supporting Ethical Phishing Campaigns. The DAMOCLES platform helps Public Administrations mitigate cybersecurity risks from human errors with its Ethical Phishing Campaign functionality. This feature allows evaluators to design and assess phishing simulations to measure user susceptibility. It provides step-by-step guidance for creating realistic phishing emails ethically. The evaluations confirm that the platform's structured design and email crafting flexibility effectively enhance cybersecurity awareness in a controlled environment.

Positive Assessment of the Usability Heuristics. The heuristic evaluation shows that DAMOCLES effectively supports ethical phishing campaigns. Its visibility of system status provides clear feedback, while user control and freedom allow evaluators to customize phishing emails through predefined options and direct text editing. The highlighting of changes in AI-generated content enhances transparency, enabling oversight of the editing process and promoting ethical practices. The platform ensures consistency, error prevention with structured input fields and clearly labeled options. This iterative AI-based authoring helps evaluators refine emails, ensuring systematic and reproducible campaigns.

Challenges and Areas for Improvement. The usability evaluation of DAMOCLES highlights improvements needed for ethical phishing campaigns. A key issue is the rigid navigation due to "Previous" and "Next" buttons, which limits users' ability to navigate freely. Adding a task overview panel could enhance usability. Additionally, the AI email authoring feature lacks visible retention of user prompts after processing, making it difficult for users to refine content. Displaying prompts alongside generated text would increase transparency. Lastly, expanding help and documentation is essential. While current inline help is basic, a task-based support system with a chatbot would better assist users in executing campaigns effectively.

Balancing Learnability, Flexibility, and Efficiency. The usability evaluation highlighted that DAMOCLES prioritizes *learnability and effectiveness* over efficiency, ensuring evaluators systematically create phishing campaigns without

skipping critical steps. As the platform evolves, enhancing efficiency through template-based automation, bulk processing, and faster navigation would benefit expert users managing multiple campaigns. The trial-and-error nature of AI-assisted phishing email creation is an intentional design choice, that promotes exploration. Unlike traditional form inputs, AI-generated content requires iterative refinement, and DAMOCLES supports this flexibility while ensuring evaluators retain control over the final output.

Control over the AI-Assisted Content. A key consideration in the design of DAMOCLES is ensuring that ethical phishing campaigns remain transparent, explainable, and controlled. The heuristic evaluation confirms strong error prevention in structured form inputs, but AI-based interactions present challenges in error recognition and recovery. Because AI-generated content is probabilistic, users may sometimes receive unexpected results that are difficult to diagnose and correct. While users can modify prompts and retry text generation, additional features such as revision history or structured AI editing suggestions could improve transparency and help users understand the reasoning behind certain AI outputs. Furthermore, to ensure that AI-generated phishing e-mails do not contain malicious content in practical applications, solutions such as Retrieval-Augmented Generation (RAG) will be explored. RAG optimizes LLM output by referring to an authoritative knowledge base outside its training data sources before generating a response. In other words, RAG extends the already advanced capabilities of LLMs to specific domains without re-training the model. It is a cost-effective approach to improve LLM output to remain relevant, accurate and useful in various contexts. In our case, an RAG for generating safe phishing emails could be built by training it on non-dangerous phishing examples.

7 Conclusion and Future Work

This paper introduced an EUD approach to simplify prompt engineering in generative AI, applied within the DAMOCLES platform for ethical phishing campaign design. By integrating example-based interactions and live programming techniques, the system enhances transparency and user control over AI-generated phishing emails. The usability evaluation confirmed that DAMOCLES effectively supports cybersecurity training, providing structured guidance and flexibility while ensuring accessibility for non-technical users.

While DAMOCLES proves effective, some areas require improvement. Navigation flexibility is limited due to its sequential workflow, making task navigation cumbersome for experienced users. Implementing a task overview panel or direct navigation options could improve usability. Additionally, AI-generated content transparency should be enhanced by retaining user prompts visibly and offering revision history or structured editing suggestions to clarify AI modifications.

As future work, we planned an update of DAMOCLES considering the emerged usability issues, and then a more formal user study with real users. This will allow us to receive more significant insights on how to improve the interaction paradigm and the user interface.

Acknowledgments. This research is partially funded by the Italian Ministry of University and Research (MUR) and by the European Union - NextGenerationEU, Mission 4, Component 2, Investment 1.1, under grant PRIN 2022 PNRR "DAMOCLES: Detection And Mitigation Of Cyber attacks that exploit human vuLnerabilitiES" (Grant P2022FXP5B) - CUP: H53D23008140001.

References

1. IFTTT. https://ifttt.com/. Accessed 14 Mar 2025
2. The most disruptive trend of 2021: No code/low code. https://www.forbes.com/sites/betsyatkins/2020/11/24/the-most-disruptive-trend-of-2021-no-code--low-code/. Accessed 14 Mar 2025
3. Akbarian, F., Fitzgerald, E., Kihl, M.: Intrusion detection in digital twins for industrial control systems. In: Proceedings of International Conference on Software, Telecommunications and Computer Networks, pp. 1–6 (2020)
4. Alexandrova, S., Cakmak, M., Hsiao, K., Takayama, L.: Robot programming by demonstration with interactive action visualizations. In: Proceedings of Robotics: Science and Systems (2014)
5. Alissa, K.A., et al.: An instrument to measure human behavior toward cyber security policies. In: Proceedings of 21st Saudi Computer Society National Computer Conference (NCC), pp. 1–6 (2018)
6. Arawjo, I., et al.: Chainforge: a visual toolkit for prompt engineering and LLM evaluation. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. ACM (2024). https://doi.org/10.1145/3613904.3642016
7. Ardito, C., et al.: User-defined semantics for the design of IoT systems enabling smart interactive experiences. Pers. Ubiquit. Comput. **24**(6), 781–796 (2020)
8. Artizzu, V., et al.: Defining configurable virtual reality templates for end users. Proc. ACM Hum.-Comput. Interact. **6** (2022)
9. Blackmon, M.H., Polson, P.G., Kitajima, M., Lewis, C.: Cognitive walkthrough for the web. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 463–470 (2002)
10. Breve, B., Desolda, G., Deufemia, V., Greco, F., Matera, M.: An end-user development approach to secure smart environments. In: Proceedings of 8th International Symposium on End-User Development. Lecture Notes in Computer Science, vol. 12724, pp. 36–52. Springer (2021)
11. Buono, P., Desolda, G., Greco, F., Piccinno, A.: Let warnings interrupt the interaction and explain: designing and evaluating phishing email warnings. In: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, pp. 1–6 (2023)
12. Corradini, I., Nardelli, E.: Building organizational risk culture in cyber security: the role of human factors. In: Ahram, T.Z., Nicholson, D. (eds.) Advances in Human Factors in Cybersecurity, pp. 193–202. Springer, Cham (2019)
13. Desolda, G., Aneke, J., Ardito, C., Lanzilotti, R., Costabile, M.F.: Explanations in warning dialogs to help users defend against phishing attacks. Int. J. Hum. Comput. Stud. **176**, 103056 (2023)
14. Desolda, G., Ferro, L.S., Marrella, A., Catarci, T., Costabile, M.F.: Human factors in phishing attacks: a systematic literature review. ACM Comput. Surv. (CSUR) **54**(8), 1–35 (2021)

15. Eckhart, M., Ekelhart, A.: Towards security-aware virtual environments for digital twins. In: Proceedings of the 4th ACM Workshop on Cyber-Physical System Security, CPSS 2018, pp. 61–72 (2018)
16. Faleiro, R., Pan, L., Pokhrel, S.R., Doss, R.: Digital twin for cybersecurity: towards enhancing cyber resilience. In: Xiang, W., Han, F., Phan, T.K. (eds.) Broadband Communications, Networks, and Systems, pp. 57–76. Springer, Cham (2022)
17. Fossati, A., Borroni, S., Marchione, D., Maffei, C.: The big five inventory (BFI). Eur. J. Psychol. Assess. (2011)
18. Ghiani, G., Paternò, F., Spano, L.D., Pintori, G.: An environment for end-user development of web mashups. Int. J. Hum. Comput. Stud. **87**, 38–64 (2016)
19. Glaessgen, E., Stargel, D.: The Digital Twin Paradigm for Future NASA and U.S. Air Force Vehicles. AIAA (2012)
20. Greco, F., Buono, P., Desiato, D., Desolda, G., Lanzilotti, R., Ragone, G., et al.: Unlocking the potential of simulated phishing campaigns: measuring the impact of interaction among different human factors. In: DAMOCLES: Detection and Mitigarion of Cyber Attacks that Exploit Human vuLnerabilitiES Workshop, co-located with AVI'24, p. 10 (2024)
21. Guenduez, A.A., Mettler, T., Schedler, K.: Technological frames in public administration: what do public managers think of big data? Gov. Inf. Q. **37**(1), 101406 (2020)
22. Gupta, M., Akiri, C., Aryal, K., Parker, E., Praharaj, L.: From ChatGPT to ThreatGPT: impact of generative AI in cybersecurity and privacy. IEEE Access **11**, 80218–80245 (2023)
23. Kim, B., Lee, D.Y., Kim, B.: Deterrent effects of punishment and training on insider security threats: a field experiment on phishing attacks. Behav. Inf. Technol. **39**(11), 1156–1175 (2020)
24. Koddebusch, M.: Exposing the phish: the effect of persuasion techniques in phishing e-mails. In: Proceedings of the 23rd Annual International Conference on Digital Government Research, pp. 78–87 (2022)
25. Kumaraguru, P., et al.: School of phish: a real-world evaluation of anti-phishing training. In: Proceedings of the 5th Symposium on Usable Privacy and Security (2009)
26. Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L.F., Hong, J.: Teaching Johnny not to fall for phish. ACM Trans. Internet Technol. **10**(2) (2010)
27. Lee, H., et al.: Constitutionmaker: supporting end-user alignment of chatbots via natural language feedback. In: Proceedings of the 2024 ACM Conference on Intelligent User Interfaces. ACM (2024). https://doi.org/10.1145/3640543.3645200
28. Lewis, C., Wharton, C.: Cognitive walkthroughs. In: Handbook of Human-Computer Interaction, pp. 717–732. Elsevier (1997)
29. Li, T.J.J., et al.: Appinite: a multi-modal interface for specifying data descriptions in programming by demonstration using natural language instructions. In: Proceedings of IEEE Symposium on Visual Languages and Human-Centric Computing, pp. 105–114 (2018)
30. Lieberman, H., Paternò, F., Wulf, V.: End User Development (Human-Computer Interaction Series). Springer, Heidelberg (2006)
31. Lim, I.k., Park, Y.G., Lee, J.K.: Design of security training system for individual users. Wirel. Pers. Commun. **90**(3), 1105–1120 (2016)
32. Mahdavi, F., et al.: Is it AI or is it me? Understanding users' prompt journey with text-to-image AI. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. ACM (2024). https://doi.org/10.1145/3613904.3642861

33. Masson, P., et al.: Directgpt: a direct manipulation interface to interact with large language models. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. ACM (2024). https://doi.org/10.1145/3613904.3642462
34. Modic, D., Anderson, R., Palomäki, J.: We will make you like our research: the development of a susceptibility-to-persuasion scale. PLoS ONE **13**(3), e0194119 (2018)
35. Mudassar Yamin, M., Hashmi, E., Ullah, M., Katt, B.: Applications of LLMs for generating cyber security exercise scenarios. IEEE Access **12**, 143806–143822 (2024)
36. Myers, B., Hudson, S.E., Pausch, R.: Past, present, and future of user interface software tools. ACM Trans. Comput.-Hum. Interact. **7**(1), 3–28 (2000)
37. Ndibwile, J.D., Kadobayashi, Y., Fall, D.: Unphishme: phishing attack detection by deceptive login simulation through an Android mobile app. In: 2017 12th Asia Joint Conference on Information Security (AsiaJCIS), pp. 38–47 (2017)
38. Nielsen, J.: 10 usability heuristics for user interface design (1994). https://www.nngroup.com/articles/ten-usability-heuristics/. Accessed 12 Mar 2025
39. Ovelgönne, M., Dumitraş, T., Prakash, B.A., Subrahmanian, V.S., Wang, B.: Understanding the relationship between human behavior and susceptibility to cyber attacks: a data-driven approach. ACM Trans. Intell. Syst. Technol. **8**(4) (2017)
40. Petrides, K.V., Furnham, A.: Trait emotional intelligence: behavioural validation in two studies of emotion recognition and reactivity to mood induction. Eur. J. Pers. **17**(1), 39–57 (2003)
41. Rizzoni, F., Magalini, S., Casaroli, A., Mari, P., Dixon, M., Coventry, L.: Phishing simulation exercise in a large hospital: a case study. Digit. Health **8**, 20552076221081716 (2022)
42. Rossetti, G., et al.: Y social: an LLM-powered social media digital twin. arXiv preprint arXiv:2408.00818 (2024)
43. Sheng, S., et al.: 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 (2007)
44. Spithoven, R., Drenth, A.: Who will take the bait? Using an embedded, experimental study to chart organization-specific phishing risk profiles and the effect of a voluntary microlearning among employees of a Dutch municipality. J. Cybersecur. **10**(1), tyae010 (2024)
45. Vaithilingam, P., et al.: Dynavis: dynamically synthesized UI widgets for visualization editing. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. ACM (2024). https://doi.org/10.1145/3613904.3642639
46. Wen, Z.A., Lin, Z., Chen, R., Andersen, E.: What.hack: engaging anti-phishing training through a role-playing phishing simulation game. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–12 (2019)
47. Workman, M.D., Luévanos, J.A., Mai, B.: A study of cybersecurity education using a present-test-practice-assess model. IEEE Trans. Educ. **65**(1), 40–45 (2022)
48. Xiong, A., Proctor, R.W., Yang, W., Li, N.: Embedding training within warnings improves skills of identifying phishing webpages. Hum. Factors **61**(4), 577–595 (2019)
49. Zamfirescu-Pereira, A., et al.: Why johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. ACM (2023). https://doi.org/10.1145/3544548.3581388