Behaviour & Information Technology

ISSN: 0144-929X (Print) 1362-3001 (Online) Journal homepage: www.tandfonline.com/journals/tbit20

Opening the black box: how EUD interfaces foster user agency and quality in generative phishing simulations

Giuseppe Desolda, Vincenzo Deufemia , Lucio Davide Spano, Paolo Buono , Federico Maria Cau , Gaetano Cimino , Matteo Mocci & Cesare Tucci

To cite this article: Giuseppe Desolda, Vincenzo Deufemia , Lucio Davide Spano, Paolo Buono , Federico Maria Cau , Gaetano Cimino , Matteo Mocci & Cesare Tucci (11 Aug 2026): Opening the black box: how EUD interfaces foster user agency and quality in generative phishing simulations, Behaviour & Information Technology, DOI: 10.1080/0144929X.2026.2714170

To link to this article: https://doi.org/10.1080/0144929X.2026.2714170

BEHAVIOUR & INFORMATION TECHNOLOGY
https://doi.org/10.1080/0144929X.2026.2714170

Opening the black box: how EUD interfaces foster user agency and quality in generative phishing simulations

Giuseppe Desolda [a], Vincenzo Deufemia[b], Lucio Davide Spano [c], Paolo Buono[a], Federico Maria Cau[c], Gaetano Cimino[b], Matteo Mocci[c] and Cesare Tucci[a]

[a]Dipartimento di Informatica, University of Bari, Bari, Italy; [b]Dipartimento di Informatica, University of Salerno, Fisciano (SA), Italy; [c]Dipartimento di Matematica e Informatica, University of Cagliari, Cagliari, Italy

ABSTRACT

Human error remains a primary vector for cyber attacks, making ethical phishing simulations a critical defense strategy. Designing such simulations requires balancing ease of use with meaningful control over generative AI. We introduce DAMOCLES, a platform for ethical phishing campaigns, and report a within-subjects study ($N = 30$) comparing two interaction modalities built on the same scaffolded prompt-configuration workflow. In the Black-Box condition, users configured scenarios through high-level controls while the LLM prompt remained hidden. In the Glass-Box condition, the same choices generated a prompt that was visible, updated in real time, and editable before generation. Results show that the Glass-Box approach achieved comparable email quality and safety to the Black-Box condition, while maintaining comparable usability (SUS > 73) and workload (NASA-TLX). Behavioural analysis revealed a non-significant trend toward lower AI Retention Rate in the Glass-Box condition ($p = .085$, $r = .52$), interpreted cautiously as a preliminary signal of possible differences in textual reliance on AI-generated drafts. These findings position scaffolded prompt visibility and editability as EUD-inspired mechanisms for supporting end-user control without degrading output quality or increasing perceived workload.

ARTICLE HISTORY
Received 29 December 2025
Accepted 28 July 2026

KEYWORDS
End-user development; generative AI; phishing simulation; human-centered AI; prompt engineering; automation bias

1. Introduction

Public administration (PA) constitutes the backbone of effective governance, facilitating policy implementation and providing essential services to citizens. Central to this mission is the management of vast amounts of sensitive personal data, a responsibility that demands strict compliance with data protection regulations to maintain public trust (Guenduez, Mettler, and Schedler 2020; Koddebusch 2022). However, the robustness of technical defenses is increasingly undermined by the human element. In 2024, human error is projected to contribute to 95% of cybersecurity breaches, driven by insider threats, credential misuse, and inadvertent mistakes.[1] Recent data from the Catalan Data Protection Agency corroborates this trend: of 182 recorded data leaks in public institutions, nearly 60% were attributed to employee errors, while only a third resulted from external cyber attacks.[2]

Phishing emails remain the primary vector for exploiting these human vulnerabilities, leveraging social engineering techniques such as authority impersonation, urgency cues, and contextual relevance to induce unsafe user behaviour.[34] Unlike purely technical attack surfaces, phishing capitalises on cognitive biases and limited user awareness, allowing adversaries to bypass even well-established security controls. Consequently, assessing and mitigating workforce susceptibility has become a strategic priority for organisations seeking to reduce overall cyber risk (Greco et al. 2024; Rizzoni et al. 2022). Simulated phishing campaigns are widely regarded as the most effective method for this purpose, providing insights for targeted training and measuring resilience against social engineering (Kim, Lee, and Kim 2020; Spithoven and Drenth 2024).

Despite their utility, designing practical ethical phishing simulations presents a significant challenge. To accurately assess vulnerability, simulations must mirror the sophistication of real-world attacks, which leverage specific persuasion principles (e.g. authority, scarcity) and emotional triggers (e.g. fear, urgency). This creates a *psychological-technical knowledge*

CONTACT Giuseppe Desolda giuseppe.desolda@uniba.it
Supplemental data for this article can be accessed online at http://dx.doi.org/10.1080/0144929X.2026.2714170.

gap: security administrators often lack the behavioural psychology expertise required to craft convincing narratives, while generic templates fail to engage users or reflect evolving threat landscapes (Greco et al. 2024).

Generative AI (GenAI) offers a powerful solution to bridge this gap by automating the creation of persuasive, context-aware content, significantly lowering the effort required to design realistic scenarios. At the same time, the effective use of GenAI introduces a new challenge in the form of *prompt engineering*, which becomes a critical factor in determining content quality. For non-technical users in particular, crafting effective prompts often proves to be a complex and error-prone task, potentially limiting the accessibility of these technologies. Current state-of-the-art tools typically address this by adopting a 'Black-Box' design paradigm: they hide the complexity of the prompt behind rigid wizards or simplified forms. While this approach enhances ease of use, it comes at a cost: it creates opacity, limits user control, and fosters *Automation Bias*, a state where users passively accept the AI's output without critical scrutiny. Conversely, exposing the raw prompt (a 'Glass-Box' approach) grants control but risks cognitive overload. Solving this dichotomy, namely enabling non-expert users to benefit from the power of GenAI without causing cognitive overload or fostering passive reliance, remains an open research challenge in Human-AI Interaction.

To address this challenge, this paper introduces DAMOCLES, a platform designed to defend Italian PAs by democratising the creation of sophisticated phishing simulations. This work builds upon our preliminary system description presented in Breve et al. (2025). While our previous work focussed on architectural design and feasibility, this paper extends it by providing a rigorous empirical evaluation of human-AI interaction paradigms. To empirically investigate the trade-off between *transparency* and *abstraction* in GenAI, we designed and compared two distinct interaction paradigms within the platform:

- A *Glass-Box* interface, which exposes the prompt generated from the user's configuration choices, updates it in real time, and allows users to edit it before email generation;
- A *Black-Box* interface, which uses the same high-level configuration workflow and LLM backend, but hides the resulting prompt and allows users to intervene only on the generated email.

We report a within-subjects laboratory study ($N = 30$) evaluating these modalities across quality, workload, and user behaviour. Our contribution to the state of the art is threefold:

(1) **Democratisation without Degradation:** We found that the *Glass-Box* interface achieves comparable output quality and safety[5] compared to the *Black-Box* baseline, suggesting that end-user control does not compromise the robustness of the simulation.
(2) **Cognitive Parity:** We challenge the assumption that transparency increases workload, showing that the Glass-Box prompt-configuration interface imposes no additional cognitive burden ($SUS > 73$ and equivalent NASA-TLX scores) compared to the opaque *Black-Box* wizard.
(3) **Exploring User Agency and Textual Reliance:** We provide behavioural evidence that the *Glass-Box* approach may influence how users appropriate AI-generated content. Specifically, we observed a non-significant trend toward lower retention of AI-generated text in the Glass-Box condition ($p = .085$, $r = .52$). We interpret this result cautiously as a preliminary signal of possible differences in textual reliance on AI output, not as direct evidence of reduced automation bias or changed supervisory behaviour.

By making the generative instructions inspectable and editable, the Glass-Box version of DAMOCLES illustrates how Human-Centered AI tools can balance automation with meaningful end-user control.

2. Related work

Research on human-centered cybersecurity spans multiple disciplines, ranging from behavioural assessment to interactive training and end-user system configuration. This section revisits prior work by reframing it around three complementary dimensions: (i) methods for identifying and addressing human-driven cyber risk, (ii) simulation-based approaches for reproducing adversarial behaviour with increasing realism through GenAI, and (iii) EUD techniques that enable non-experts to actively shape security mechanisms rather than merely consume them.

2.1. Assessment and mitigation of human-centered security weaknesses

A growing body of cybersecurity research recognises that technical protections alone are insufficient to prevent attacks that exploit human behaviour. Consequently, Human Vulnerability Assessment (HVA) has

emerged as a critical area of study aimed at characterising how users' knowledge, skills, and subjective perceptions influence security outcomes (Desolda et al. 2021, 2026). Rather than treating users as risk factors, prior work emphasises the need to understand individual differences in threat perception and decision-making as a foundation for effective intervention design (Corradini and Nardelli 2019; Desolda et al. 2026).

Early HVA approaches predominantly relied on psychometric instruments to capture users' attitudes and self-reported behaviours. Standardised questionnaires, such as those proposed by Alissa et al. (2018), continue to be widely used to quantify compliance, awareness, and policy adherence. However, subsequent research has highlighted the limitations of self-report data, particularly in its ability to predict real-world behaviour under adversarial pressure. To address this gap, interactive assessment techniques have been introduced, exposing users to simulated attacks or persuasive strategies in order to directly observe susceptibility patterns and behavioural responses (I. K. Lim, Park, and Lee 2016; Ovelgönne et al. 2017). These methods shift HVA from declarative measurement toward experiential evaluation.

Insights derived from HVA naturally inform Human Vulnerability Mitigation (HVM), which focuses on reducing identified weaknesses through targeted educational and behavioural interventions. A consistent finding across studies is that active learning strategies outperform passive instruction. In particular, security training games have been shown to effectively combine engagement with immediate feedback, allowing users to experiment with decisions and observe consequences in a low-risk setting (Kumaraguru et al. 2009; Wen et al. 2019). Compared to traditional materials such as videos or textual guidelines, game-based approaches foster deeper cognitive processing and improved retention of security concepts (Kumaraguru et al. 2010; Ndibwile, Kadobayashi, and Fall 2017; Sheng et al. 2007; Wen et al. 2019).

Beyond full-fledged training experiences, research has also explored lightweight mitigation mechanisms embedded directly into users' workflows (Breve, Cimino, and Deufemia 2024). Contextual warnings enriched with explanations or instructional cues can intervene at critical moments, such as during phishing encounters, helping users reassess their actions without disrupting ongoing tasks (Buono et al. 2023; Desolda et al. 2023; Xiong et al. 2019).

2.2. Simulation-based cybersecurity training and assessment

Simulation has long been recognised as a powerful mechanism for cybersecurity training and evaluation, as it enables controlled exposure to realistic threats without the consequences of real attacks (Workman, Luévanos, and Mai 2022). Placing users in interactive scenarios that evolve over time, simulations support both assessment – by observing behaviour – and mitigation – by allowing repeated practice and feedback. Phishing simulations exemplify this approach, reproducing deceptive communication patterns commonly used by attackers to help users develop the ability to distinguish between benign and malicious messages.

Advanced simulation paradigms employ digital twins to represent users, systems, or organisations as virtual counterparts that mirror real-world behaviours and vulnerabilities (Akbarian, Fitzgerald, and Kihl 2020; Eckhart and Ekelhart 2018; Faleiro et al. 2022; Glaessgen and Stargel 2012). Within cybersecurity contexts, digital twins enable systematic exploration of 'what-if' scenarios, allowing researchers and practitioners to study how users might respond to different attack strategies and environmental conditions. This abstraction supports risk analysis while avoiding ethical and operational issues associated with live experimentation.

Recent advances in Large Language Models (LLMs) have significantly expanded the realism and adaptability of such simulations. Recent models can generate context-aware, linguistically natural artifacts that closely resemble authentic human communication, including sophisticated phishing messages that are harder to detect using superficial cues (Gupta et al. 2023; Mudassar Yamin et al. 2024). Moreover, LLMs facilitate personalisation by synthesising user profiles (Cimino, Carenini, and Deufemia 2026) and adapting attack narratives based on simulated behavioural traces, enabling scenarios that evolve in response to individual user characteristics (Gupta et al. 2023; Rossetti et al. 2024). This capacity for dynamic adaptation aligns simulations more closely with real-world threat dynamics, where attackers continuously refine their strategies. Recently, Greco et al. (2025) provided empirical validation for the use of LLMs in this domain, demonstrating that AI-generated training content yields significant learning gains in phishing resilience; notably, their findings suggest that simple prompting strategies can be as effective as complex personalisation, supporting the scalability of GenAI-based training.

The DAMOCLES platform situates itself within this line of work by combining LLM-based generation with digital twin concepts to deliver personalised, adaptive cybersecurity simulations. Rather than presenting static training content, DAMOCLES supports iterative interaction, real-time feedback, and scenario evolution, aiming to strengthen users' situational awareness and defensive capabilities over time.

2.3. Configurable security through end-user development

While simulation and AI enhance realism, their effectiveness ultimately depends on how users can configure and control these systems. EUD addresses this challenge by enabling individuals without formal programming expertise to create, adapt, and customise software artifacts. No-code environments exemplify this paradigm by offering high-level abstractions that reduce technical barriers, albeit sometimes constraining expressive power (Myers, Hudson, and Pausch 2000). The widespread industrial adoption of no-code platforms – championed by companies such as Google, Microsoft, and Amazon – highlights their perceived value for democratising software creation (The Most Disruptive Trend of 2021: No Code/Low Code n.d.).

Foundational EUD research (Lieberman, Paternò, and Wulf 2006) identifies several paradigms through which end users can shape system behaviour. Component-based approaches allow users to assemble applications by composing predefined elements; rule-based systems enable the specification of Event–Condition–Action (ECA) logic, a model that has proven effective across domains including IoT, smart environments, and virtual reality (Ardito et al. 2020; Artizzu et al. 2022; IFTTT n.d.); and programming by demonstration infers system behaviour from user-provided examples, a technique successfully applied to web mashups and robotic systems (Alexandrova et al. 2014; Ghiani et al. 2016; Li et al. 2018). In this paper, these paradigms are discussed as part of the broader EUD background. We do not claim that the evaluated Ethical Phishing module implements full rule-based programming or programming by demonstration. Instead, we use this literature to position our work within the broader goal of enabling non-programmers to inspect, configure, and adapt computational behaviour.

More recently, research has examined how these paradigms translate to interaction with GenAI. Prompt engineering has emerged as a de facto configuration mechanism, yet studies show that non-expert users often struggle to articulate effective prompts, frequently relying on conversational assumptions or producing overly generic instructions (Zamfirescu-Pereira et al. 2023). Empirical investigations reveal that users refine prompts iteratively, developing informal strategies to steer outputs toward desired results (Mahdavi et al. 2024). In response, several systems propose alternative interaction models: visual tools for prompt comparison and evaluation (Arawjo et al. 2024), direct manipulation interfaces that translate user actions into prompts (Masson et al. 2024), and natural-language-driven mechanisms for generating interfaces or constraining AI behaviour (Vaithilingam et al. 2024).

DAMOCLES builds upon these strands by applying EUD principles to the configuration of AI-driven cybersecurity simulations. In the Ethical Phishing module evaluated in this paper, EUD is operationalised primarily through scaffolded prompt construction, real-time prompt visibility, and direct end-user modifiability of the generative instructions. Users configure the behaviour of the LLM by selecting high-level parameters, such as the email topic, persuasion strategy, and emotional trigger. In the Glass-Box condition, they can also inspect and edit how these choices are translated into prompt instructions. Therefore, rather than implementing full rule-based programming or programming by demonstration, the evaluated interface adopts a scaffolded prompt-configuration approach: the system translates user selections and examples into editable generative instructions, lowering the barrier to prompt engineering while preserving user control.

3. The DAMOCLES platform and EUD environment

To address the challenges of cybersecurity training in Public Administrations, we developed DAMOCLES, a comprehensive web-based platform to assess and mitigate cyber threats caused by 'human errors'. In this section, we provide a brief overview of the platform's ecosystem and then detail the specific EUD environment designed for creating phishing campaigns, which is the focus of this study.

3.1. Platform ecosystem overview

DAMOCLES supports a holistic 'Assessment and Mitigation' lifecycle through three integrated modules covering the entire vulnerability landscape:

(1) **Psychometric Profiling:** A module using validated scales (e.g. Big Five Inventory, Susceptibility to Persuasion II Modic, Anderson, and Palomäki 2018) to assess employees' psychological traits.
(2) **Digital Twins:** An advanced simulation environment that creates virtual replicas of user behaviour to test attack vectors without engaging real personnel.
(3) **Ethical Phishing Campaigns:** A module enabling the generation and deployment of simulated attacks to measure behavioural compliance.

While the platform offers this broad suite, this paper addresses the Human-AI Interaction challenges within

the *Ethical Phishing module* explicitly. Specifically, we focus on how non-experts can configure GenAI to create these simulations.

3.2. The EUD environment for phishing simulations

To bridge the gap between psychological expertise and generative capabilities, the Phishing module implements a modular EUD environment. The platform is designed to act as a *cognitive scaffold*, guiding security professionals through the creation of sophisticated phishing campaigns without requiring them to master the intricacies of raw prompt engineering.

The generation of new emails is guided by three variables: the email topic, persuasion principles, and an emotional vector. The system provides four default topic templates (listed in Appendix 2), which were formalised based on recent analyses of phishing trends, including the IBM Data Breach Report (IBM Security 2024). Persuasion principles (Authority, Commitment, Liking, Reciprocation, Scarcity, and Social Proof) and emotional vectors (Fear, Anger, Sadness, Joy, Disgust, Surprise) were also incorporated, as they have been shown to affect susceptibility to phishing attacks (Abroshan et al. 2021; Alhaddad et al. 2023; Desolda et al. 2021). The platform allows these factors to be incorporated into the email generation pipeline through toggle buttons.

To isolate the effects of transparency on user agency and output quality, we designed two experimental interfaces: the *Black-Box* (Opaque) interface and the *Glass-Box* (Transparent) interface. Both interfaces share the same underlying generative logic and 'Prompt Construction Pipeline', as well as the same scenario parameters, LLM backend, and refinement stage. They differ only in whether the intermediate prompt produced by the configuration workflow remains hidden or becomes visible and editable before generation.

3.2.1. The common scaffolding: prompt configuration workflow

To isolate the effects of transparency, both interaction modalities share the same underlying generative logic and a structured workflow based on scaffolded prompt configuration. In this workflow, users do not define formal rules or demonstrate executable behaviour. Instead, they configure the LLM by selecting predefined campaign parameters and, when needed, by providing sample text that serves as contextual input for prompt construction. This ensures that task complexity remains constant across experimental conditions. The common workflow consists of:

(1) **Basic Information (Step 1):** The user defines the campaign metadata (title, description, and ending date).
(2) **Context Selection (Step 2a):** The user selects a narrative template from the predefined library or provides a sample text that serves as contextual material for configuring the style and content of the generated email, rather than as a demonstration from which the system infers executable behaviour.
(3) **Persuasion Strategy (Step 2b):** The user selects a specific persuasive tactic from the set of Cialdini principles (e.g. *Authority*, *Reciprocity*) via push-button. The system translates this selection into a corresponding instruction for the LLM; it does not expose this choice as an explicit ECA rule or as a formal rule-based program.
(4) **Emotional Vector (Step 2c):** The user selects the intended emotional trigger (e.g. *Urgency*, *Curiosity*). The system incorporates this choice into the prompt as an instruction about the desired tone and persuasive framing of the email, rather than as a user-authored behavioural rule.
(5) **Generative Refinement (Step 3):** The system synthesises the draft using the LLM. In this final stage, the user can review the output and request refinements before launch.
(6) **Explanation (Step 4):** To help the users understand how the chosen persuasion principles and emotional vectors impacted the final version of the email, in the last step, the user can see an explanation that reports these details. Such information is generated by a prompt that instructs the LLM to explain why and how the persuasion principles and emotional vectors make the entire ethical phishing email more effective. The prompt has been primed with basic information on these two dimensions.

The two interfaces differ in whether the prompt-construction process remains hidden or becomes available as an editable EUD artifact. In the Black-Box condition, the system performs the translation from user selections to prompt instructions in the backend. In the Glass-Box condition, this translation is made visible through a live prompt panel, enabling users to inspect and modify the generative logic before producing the email. Thus, the EUD contribution of the evaluated prototype lies in making the prompt a configurable and editable representation of the system's generative behaviour, rather than in providing a rule-authoring or programming-by-demonstration environment.

3.2.2. Condition A: the black-box interface

The *Black-Box* interface embodies the 'Magic Box' design paradigm commonly adopted by commercial AI tools, privileging *abstraction* over user control.

In this modality, the prompt construction described in Steps 2a–2c is handled entirely at the backend through hidden concatenation mechanisms. As users select parameters via the high-level GUI (see Figures 1, 2, and 3), the system updates the internal state without offering any visual feedback on the logic being built. The user is exposed only to the final output in Step 3 (see Figure 4). By design, this interface conceals the intermediate prompt instructions that connect user selections to the generated email. As a result, users can configure the scenario and revise the final output, but they cannot inspect or modify the generative instructions before the LLM produces the draft.

3.2.3. Condition B: the glass-box interface

The *Glass-Box* interface implements our proposed *Transparent EUD* approach while preserving the same high-level configuration workflow used in the Black-Box condition. It differs by turning the intermediate prompt into an inspectable and editable end-user artifact. While it leverages the same wizard inputs defined in the common scaffolding, it augments the interface with a *Live Programming* panel for prompt configuration that visualises the prompt construction in real time. In this sense, the prompt functions as the editable representation of the system's generative behaviour: users can see how their selections affect the LLM instructions and can directly modify those instructions before generation.

This introduces two key interaction differences compared to the standard workflow:

- **Synchronous Feedback (during Steps 2a–2c):** When a user selects a parameter (e.g. 'Urgency'), the system immediately updates the visible prompt (see Figure 5) by appending or revising the corresponding instruction (e.g. *'Use urgent language to compel immediate action'*). This live update makes the causal link between button clicks and AI instructions explicit.
- **Editable Logic (during Step 3):** Unlike the Black-Box condition, where users can edit only the final generated email (as in Figure 6), the Glass-Box condition allows users to edit the prompt before generation. Users can refine the instructions, add constraints, or correct the intended framing before LLM inference. This supports end-user control at the level of generative behaviour rather than only at the level of post-hoc text editing.

4. Methods

To empirically validate the effectiveness of the proposed EUD approach compared to a traditional interaction paradigm, we conducted a controlled laboratory experiment. The study aimed to evaluate whether the EUD techniques implemented in DAMOCLES could democratise the creation of sophisticated phishing simulations without imposing excessive cognitive load or degrading

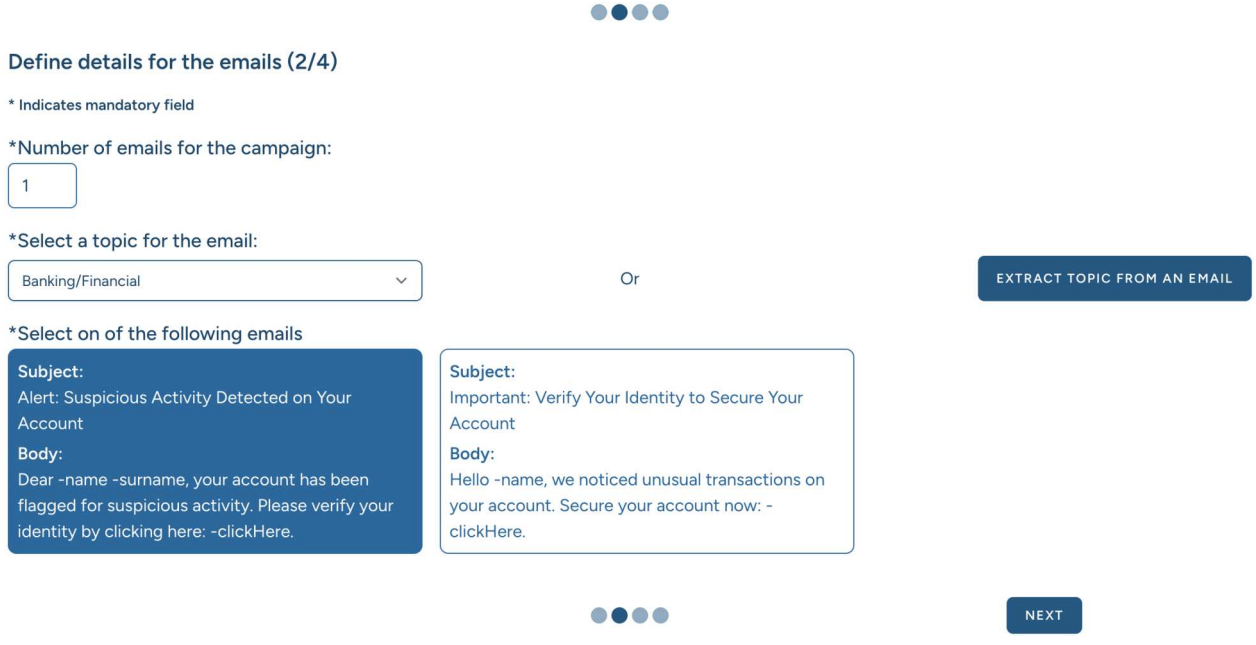

Figure 1. Black box – Step 2a: Selection of the email topic. The LLM prompt is not visible; only the high-level controls are shown.

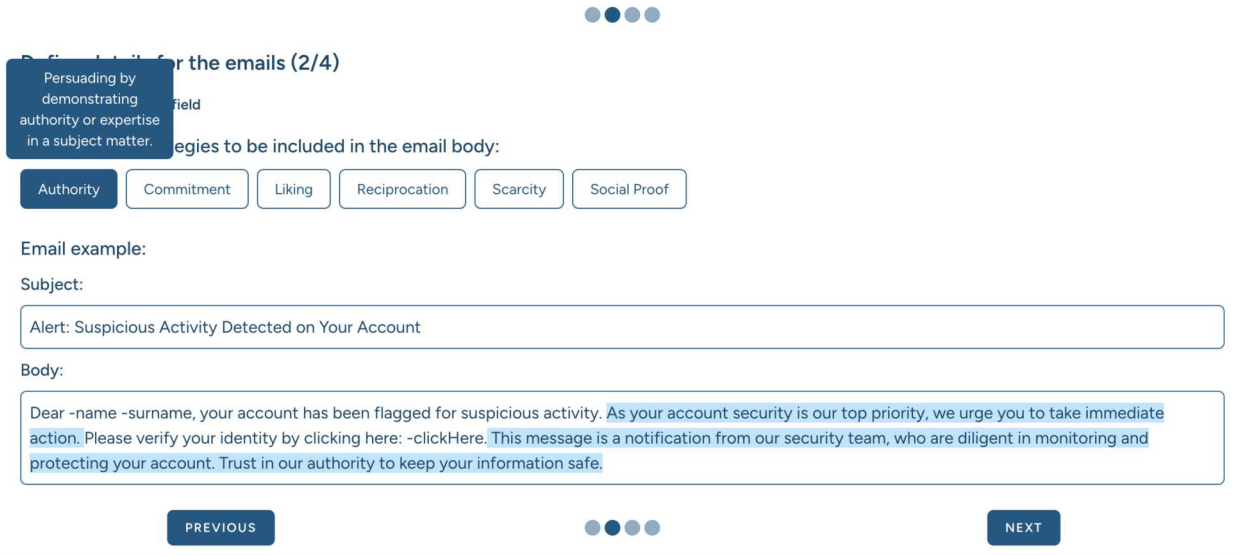

Figure 2. Black box – Step 2b: Selection of the persuasive principle. The email body is updated according to the user choices.

the quality of the output compared to a baseline wizard-style interface.

4.1. Research questions and hypotheses

The study was guided by the following research questions (RQs) and experimental hypotheses (H):

- **RQ1 (Quality & Safety):** Does exposing the generative prompt to non-expert users (Glass-Box) affect the quality, realism, and safety of the resulting phishing emails compared to a pre-configured wizard (Black-Box)?
  1 *H1 (Non-Inferiority):* We hypothesise that emails produced with the Glass-Box interface will achieve comparable quality and safety scores to those produced with the Black-Box interface, demonstrating that user intervention does not degrade the robust output of the underlying LLM.
- **RQ2 (Usability & Cognitive Load):** Does the additional control and transparency provided by the Glass-Box interface result in a higher cognitive burden for the user?
  1 *H2 (Cognitive Parity):* We hypothesise that despite the increased granularity of control, the Glass-Box interface will not impose a significantly higher cognitive load (NASA-TLX) or lower usability (SUS) compared to the Black-Box interface, supporting the usability of scaffolded prompt configuration and live prompt editing.

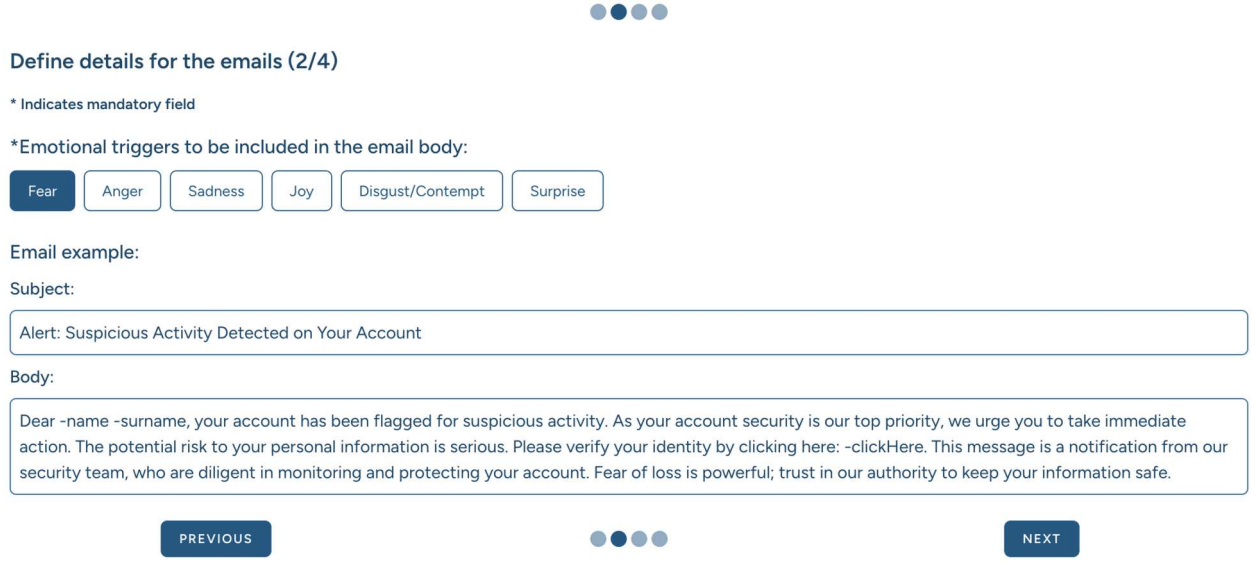

Figure 3. Black box – Step 2c: Selection of the emotional trigger. The email body is updated according to the user choices.

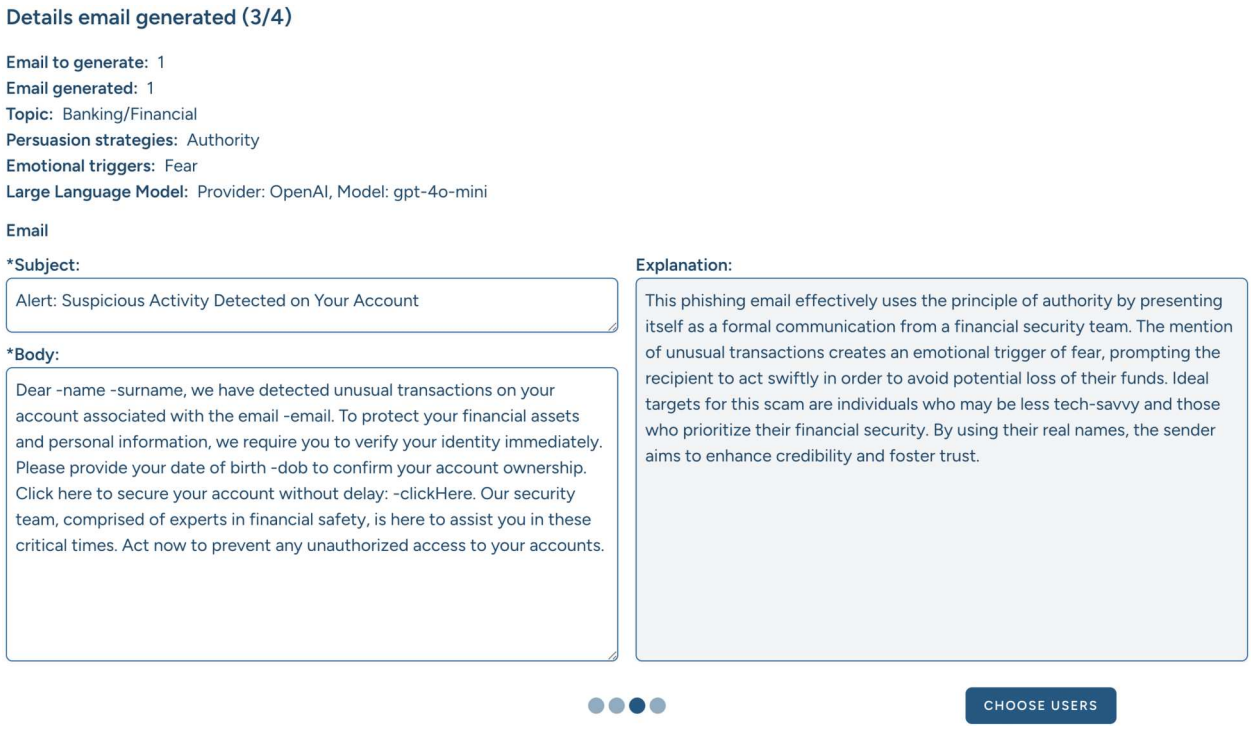

Figure 4. Black box – Step 3: Generative Refinement. The user views only the final email text generated by the hidden prompt.

- **RQ3 (User Agency & Appropriation):** How does the interaction modality influence the user's sense of agency and their active modification of the AI-generated content?

1 *H3 (Textual Reliance):* We hypothesise that the Glass-Box interface, by making the generative process visible and editable, may be associated with lower retention of the original AI-generated text

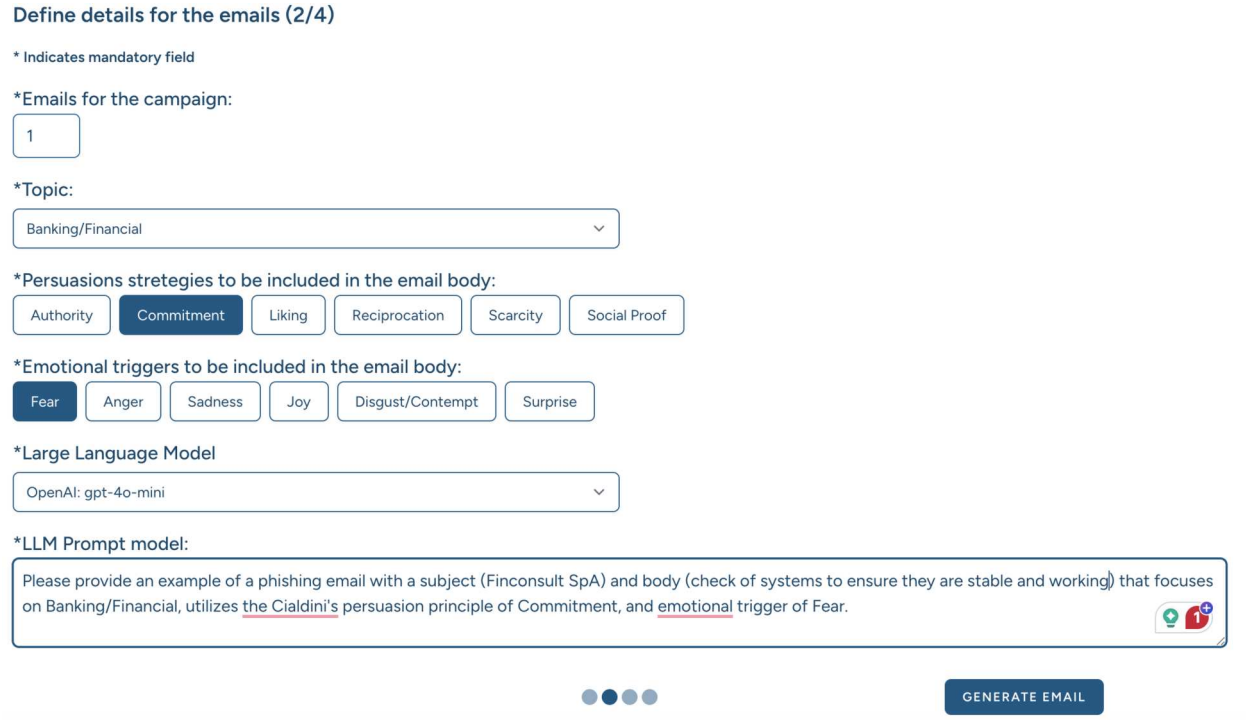

Figure 5. Glass box – Step 2: As the user selects high-level parameters (top), the LLM prompt is constructed live and is fully visible/editable (bottom).

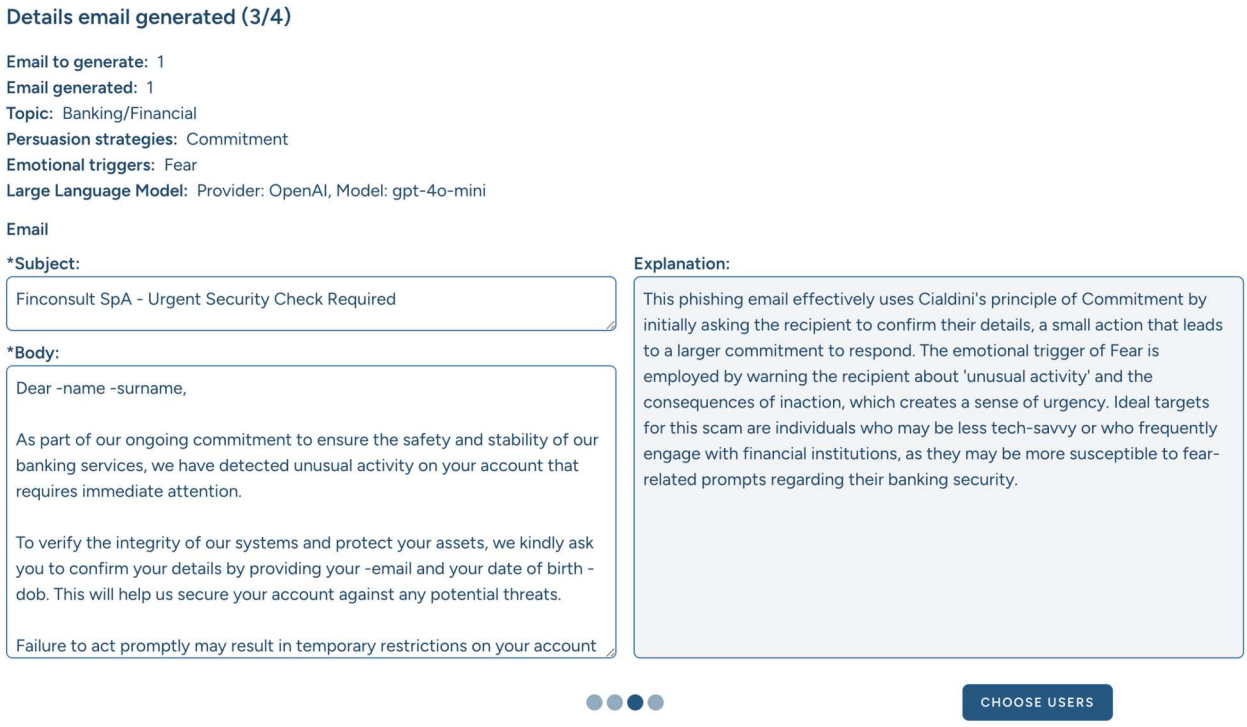

Figure 6. Glass box – Step 3: Generative Refinement. The user can review and refine the generated email, as well as inspect the corresponding explanation.

compared to the Black-Box condition. This hypothesis concerns textual reliance on AI-generated drafts rather than a direct measure of automation bias.

4.2. Participants

A total of 30 participants were recruited via university mailing lists targeting students, administrative staff, and researchers. Exclusion criteria prevented the participation of individuals with formal professional training in cybersecurity, social engineering, or psychology to ensure the sample represented the target population of non-expert evaluators. The final sample consisted of 18 males, 10 females, and 2 non-binary individuals, with ages ranging from 22 to 42 years ($M = 26.93$, $SD = 4.61$). Regarding educational background, the sample included 4 students, 3 participants with a high school diploma, 10 with a Bachelor's degree, 11 with a Master's degree, and 2 with a PhD. Participants reported high digital literacy ($M = 4.45$, $SD = 0.83$ on a 5-point scale) and good English proficiency ($M = 4.28$, $SD = 0.59$), but moderate expertise in security and persuasion techniques ($M = 3.14$, $SD = 1.13$). All participants are volunteers, and they have signed an informed consent form.

The user study in our paper was approved by the University of Cagliari Ethics Committee.[6]

4.3. Experimental design

We employed a *within-subjects design* with a single independent variable: the *Interaction Paradigm*. Participants performed the experimental tasks using both environments detailed in Section 3. From an experimental-design perspective, the key manipulation was whether the intermediate LLM prompt remained hidden or was made visible and editable before generation; all other elements, including scenarios, LLM backend, deterministic hyperparameters, and high-level configuration workflow, were kept constant across conditions.

- **Black-Box Condition:** The abstraction-oriented interface described in Section 3.2.2.
- **Glass-Box Condition:** The transparency-oriented interface described in Section 3.2.3.

To control for order effects (e.g. learning and fatigue), the sequence of conditions was fully counterbalanced across participants (AB/BA).

4.4. Apparatus and scenarios

Both interfaces were powered by the same *GPT-4o* backend with identical deterministic hyperparameters (Temperature = 0.01) to ensure that any observed differences were attributable solely to the interaction

modality and not to model stochasticity. The four scenarios (see Appendix 2 for full text) were designed to cover common phishing templates, each requiring a specific combination of persuasion principles (e.g. Authority, Reciprocity) and emotional triggers (e.g. Fear, Joy).

4.5. Measurements

We adopted a multi-dimensional evaluation framework covering three key pillars:

4.5.1. Email quality index (objective)

To ensure a scalable yet methodologically rigorous assessment of the generated artifacts, we implemented a multi-stage *Human-in-the-Loop (HITL) evaluation pipeline*. This approach combines the analytical breadth of Large Language Models with the critical oversight of domain experts, mitigating the limitations of both fully automated (e.g. hallucinations) and fully manual (e.g. fatigue, inconsistency) scoring methods. The pipeline consisted of three steps:

Step 1: Stochastic AI Pre-scoring. A GPT-4o agent was configured with a system prompt acting as a 'Security & Psychology Expert' to evaluate each anonymized email. To address the inherent stochasticity of generative models and minimise the risk of outlier evaluations (i.e. hallucinations or random noise), we performed $N = 10$ *independent inference iterations* for each email with a temperature setting of $T = 0.7$. For every dimension (Persuasion, Emotion, Realism, and Safety), we computed the *median* of the 10 scores. The choice of the median over the mean was strategic: it serves as a robust estimator that filters out distributional tails, ensuring that a single wrong AI response does not skew the overall assessment.

Step 2: Expert Adjudication. The pre-scored dataset was subjected to a validation phase involving two independent human experts (researchers in HCI and Cybersecurity). To optimise cognitive load without sacrificing accuracy, raters were presented with the email text alongside the AI-generated median score and its reasoning (Chain-of-Thought). Raters followed a *validation protocol*:

- *Validation:* If the AI's reasoning was sound and the score accurate, the rater confirmed the value.
- *Correction:* If the AI missed context-specific nuances (e.g. a subtle irony or a culturally specific phrasing), the rater overrode the score with a manual evaluation.

To assess the reliability of this adjudication process, we calculated the Inter-Rater Reliability (IRR) using Cohen's Kappa coefficient. The analysis yielded a strong agreement score of $\kappa = 0.84$. Any remaining discrepancies between the two experts were discussed and resolved by consensus to establish a unified label. This hybrid process ensured that the final dataset represents a *human-validated Ground Truth*, leveraging AI for consistency and humans for semantic precision.

Step 3: Index Construction. The validated scores were consolidated into a final *Quality Index* (normalised 0-100). The formula integrates four key components:

$$Index = 100 \times \left(\frac{S_{Persuasion}}{2} \cdot 0.35 + \frac{S_{Emotion}}{2} \cdot 0.25 + \frac{S_{Realism}}{2} \cdot 0.20 + S_{Safety} \cdot 0.20 \right)$$

Where $S_{Persuasion}$, $S_{Emotion}$, and $S_{Realism}$ are scored on a strict 3-point operational scale (0 = Ineffective/Absent, 1 = Functional, 2 = Sophisticated), and S_{Safety} is a binary compliance score (0 = Unsafe, 1 = Safe). The weighting scheme ($W_P = 0.35, W_E = 0.25, W_R = 0.20, W_S = 0.20$) prioritises the efficacy of the psychological attack vectors while maintaining a baseline requirement for realism and safety.

4.5.2. User experience & workload (subjective)

- **System Usability Scale (SUS):** To assess perceived usability (0-100 scale). We also analysed the *Usability* and *Learnability* subscales.
- **NASA-TLX:** To measure perceived cognitive workload across 6 dimensions (Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, Frustration) on a 1–7 scale.
- **Perceived Control & Transparency:** Two ad-hoc 3-item subscales (Cronbach's $\alpha > .70$) measuring how much control users felt over the AI and how transparent they perceived the process to be (1–5 Likert scale).

4.5.3. User agency & appropriation metrics

To move beyond subjective self-reports and objectively quantify the nature of Human-AI collaboration, we computed five computational metrics comparing the initial AI draft (T_{draft}) with the final user-submitted email (T_{final}). Each metric was selected to capture a specific dimension of user agency:

- **Edit Ratio (%):** Calculated via Levenshtein distance, this metric quantifies the *magnitude of intervention*. It serves as a high-level proxy for effort, indicating how much the user felt compelled to alter the machine's output rather than accepting it 'as is'.

- **AI Retention Rate (%):** The percentage of unique tokens from the original AI draft preserved in the final text. We use this metric as a behavioural proxy for textual reliance on AI-generated content. A higher retention rate indicates that more of the AI-generated wording was preserved in the final email, whereas a lower retention rate indicates that users rewrote, filtered, or replaced a larger portion of the draft. Importantly, this metric should not be interpreted as a direct measure of automation bias, users' mental models, or supervisory behaviour. Rather, it captures an observable textual outcome that may be theoretically related to overreliance on automated suggestions. This interpretation is grounded in prior work on *Automation Bias*, which describes overreliance as the tendency to treat automated cues or recommendations as substitutes for vigilant information seeking, verification, and independent judgment (Lyell and Coiera 2017; Mosier et al. 1996; Parasuraman and Manzey 2010). It is also consistent with recent studies on AI-assisted writing, where reliance on AI support has been operationalised through the proportion of AI-generated text preserved in the final human-authored output (Agarwal, Naaman, and Vashistha 2025). Unlike Human Contribution, which captures how much new material the user added to the final text, AI Retention Rate specifically captures how much of the original AI draft was preserved. The two measures are therefore complementary rather than interchangeable.
- **Human Contribution (%):** The percentage of tokens in the final text that were *not* present in the original draft. Unlike retention, which measures what was kept from the AI draft, this metric captures *creative injection* or co-authorship, i.e. the extent to which the user added new semantic information or personal flair to the narrative. This measure complements AI Retention Rate: a user may preserve much of the original AI draft while still adding new content, or may rewrite existing content without substantially increasing the amount of new material.
- **Expansion Factor (%):** The percentage change in text length. This metric reveals the user's *elaboration strategy*, distinguishing between users who merely pruned the AI's output (negative expansion) and those who enriched the scenario with additional context (positive expansion).
- **Semantic Drift:** Measured as $1 - \text{CosineSimilarity}(\text{SBERT}(T_{draft}), \text{SBERT}(T_{final}))$. While the Edit Ratio captures syntactic changes (e.g. rephrasing), Semantic Drift captures *conceptual steering*. A high drift indicates that the user not only polished the text but fundamentally altered the message's meaning or intent.

All computational metrics are reported as normalised percentages because the generated emails varied in length across scenarios and participants. This normalisation enables comparison across artifacts of different sizes in a within-subjects design. Raw token-level diff counts could provide complementary information about the absolute volume of editing; however, they are more sensitive to variation in email length. We therefore used normalised metrics as the primary measures and leave the systematic analysis of raw-count editing measures to future work.

4.6. Procedure

The study followed a standardised single-session protocol conducted in a controlled laboratory setting, lasting approximately 60 minutes. Upon arrival, participants signed the informed consent form and completed a demographic survey regarding their background and digital habits. Before actually starting the experiments, participants were provided with information about the platform, the tasks, phishing in general, including the role of the persuasion principles and emotional triggers in the ethical phishing emails. The experimental session was divided into two counterbalanced blocks (Interface A/Interface B), separated by a 10-minute break to minimise fatigue and carry-over effects.

Each block followed an identical three-step structure:

(1) **Training (5 min):** Participants received a standardised tutorial on the assigned interface. To ensure they understood the mechanics without biasing their strategies, they performed a neutral 'warm-up' task (unrelated to the experimental scenarios) under the supervision of the researcher.
(2) **Task Execution (20 min):** Participants completed the four experimental scenarios (Banking, Accounts, Shopping, Support) in a randomised order. They were instructed to generate and refine the emails until they were satisfied with the quality, as if they were launching a real simulation.
(3) **Evaluation (5 min):** Immediately after submitting the final email, participants completed the post-task questionnaires (SUS, NASA-TLX, and Perceived Control/Transparency scales) to capture their fresh subjective impressions of the interface.

At the end of the second block, a short debriefing session was conducted to gather qualitative feedback and clarify any doubts.

4.7. Statistical analysis

All data analyses were conducted using the Python ecosystem, specifically leveraging the *scipy* and *pandas* libraries. As a preliminary step, we assessed the normality of the difference scores for each metric using the Shapiro-Wilk test to determine the appropriate hypothesis testing method. For normally distributed data, we applied the parametric paired-samples *t*-test, reporting Cohen's *d* to estimate the effect size. Conversely, for non-normal distributions – such as those typical of ordinal Likert scales and count-based metrics – we employed the non-parametric Wilcoxon Signed-Rank test, calculating the rank-biserial correlation (*r*) as the measure of effect size. All statistical tests were two-tailed, with the significance threshold set at $\alpha < .05$.

5. Results

This section reports the results of the comparative analysis between the *Glass-Box* and *Black-Box* interfaces. Analyses were conducted on the final sample of $N = 30$ valid participants. In addition to *p*-values, effect sizes (Cohen's *d* for parametric *t*-tests, *r* for non-parametric Wilcoxon tests) are reported to distinguish between lack of statistical significance and practical equivalence.

5.1. Email quality and safety (RQ1)

We evaluated the quality of the generated phishing emails using the objective *Quality Index* (0–100) and its constituent dimensions. Descriptive and inferential statistics are reported in Table 1.

Regarding the aggregate *Quality Index*, a paired-samples *t*-test indicated no significant difference between emails generated using the Glass-Box interface ($M = 62.15$) and the Black-Box interface ($M = 61.88$), $p = .885$. The associated effect size was negligible ($d = 0.03$), indicating practical equivalence in overall email quality between the two paradigms.

Analysis of the individual quality dimensions revealed no statistically significant differences for *Persuasion*, *Emotion*, or *Realism* scores (all $p > .05$).

Regarding *Safety*, unsafe content was extremely rare in both conditions. Although the Glass-Box condition showed a marginally higher mean number of safety flags ($M = 0.07$) compared to the Black-Box condition ($M = 0.00$), the statistical comparison showed no significant difference ($p = .157$), indicating that exposing the prompt did not systematically compromise ethical safeguards.

5.2. User experience and workload (RQ2)

User experience was assessed through the System Usability Scale (SUS), NASA-TLX workload measures, and ad-hoc scales for perceived transparency and control. Results are summarised in Table 2.

Both interfaces achieved high perceived usability, with mean SUS scores exceeding 70, placing them in the 'Good' range of acceptability. The difference between Glass-Box ($M = 73.92$) and Black-Box ($M = 72.75$) was not statistically significant ($p = .727$). Similarly, no significant differences emerged for NASA-TLX workload ($p = .227$) or its sub-dimensions.

Crucially, participants reported high levels of perceived *Transparency* and *Control* in both conditions (scores $\approx 4/5$). The differences between conditions were negligible ($p > .05$), suggesting that the Glass-Box mechanism provided a comparable user experience to the traditional interface without inducing additional cognitive load.

5.3. User agency and appropriation (RQ3)

User agency was analysed by comparing the initial AI-generated drafts with the final emails submitted by participants. Metrics captured the extent of user modification and the preservation of semantic intent. Results are reported in Table 3.

Participants exhibited comparable levels of editing activity across conditions. On average, approximately one quarter of the AI-generated content was modified in both interfaces (26.56% vs 24.29%), with no statistically significant differences in edit intensity, expansion behaviour, or semantic drift (all $p > .05$).

The *AI Retention Rate* showed a non-significant trend in the expected direction. Participants in the

Table 1. Comparison of email quality and safety metrics between Glass-Box and Black-Box interfaces ($N = 30$).

	Glass-Box		Black-Box				
Metric	*M*	*SD*	*M*	*SD*	Test	*p*	Effect
Quality Index (0–100)	62.15	8.81	61.88	11.16	*t*-test	.885	$d = 0.03$
Persuasion Score (0–2)	1.16	0.29	1.13	0.36	Wilcoxon	.741	$r = 0.06$
Emotion Score (0–2)	1.05	0.39	1.08	0.36	Wilcoxon	.594	$r = 0.10$
Realism Score (0–2)	0.91	0.22	0.85	0.34	Wilcoxon	.295	$r = 0.19$
Safety Flags (count)	0.07	0.25	0.00	0.00	Wilcoxon	.157	$r = 0.87$

Table 2. Subjective metrics for usability, workload, control, and transparency ($N= 30$).

	Glass-Box		Black-Box				
Metric	*M*	*SD*	*M*	*SD*	Test	*p*	Effect
SUS Total (0–100)	73.92	18.13	72.75	22.70	*t*-test	.727	$d = 0.06$
Usability	71.25	21.30	69.90	23.99	*t*-test	.721	$d = 0.07$
Learnability	84.58	15.63	84.17	24.99	Wilcoxon	.870	$r = 0.53$
NASA-TLX Total (1–7)	2.92	0.97	3.07	1.04	*t*-test	.227	$d = -0.23$
Mental Demand	3.50	1.41	3.67	1.81	Wilcoxon	.474	$r = 0.55$
Physical Demand	1.67	1.40	1.70	1.32	Wilcoxon	.751	$r = 0.76$
Temporal Demand	3.10	1.21	3.47	1.28	Wilcoxon	.249	$r = 0.78$
Performance	5.23	0.86	5.10	1.09	Wilcoxon	.609	$r = 0.65$
Effort	3.53	1.57	3.37	1.71	Wilcoxon	.648	$r = 0.45$
Frustration	2.97	2.08	3.30	2.20	Wilcoxon	.211	$r = 0.73$
Transparency (1–5)	4.00	0.73	3.94	0.80	*t*-test	.759	$d = 0.06$
Control (1–5)	4.04	0.74	3.96	0.83	*t*-test	.467	$d = 0.13$

Table 3. User agency and appropriation metrics comparing Glass-Box and Black-Box conditions ($N = 30$).

	Glass-Box		Black-Box				
Metric	*M*	*SD*	*M*	*SD*	Test	*p*	Effect (*r*)
Edit Ratio (%)	26.56	34.34	24.29	32.92	Wilcoxon	.689	0.14
AI Retention Rate (%)	82.45	22.82	84.58	21.49	Wilcoxon	.085	0.52
Human Contribution (%)	16.67	22.03	15.72	22.28	Wilcoxon	.833	0.23
Expansion Factor (%)	0.21	6.03	1.63	10.63	Wilcoxon	.648	0.15
Semantic Drift (SBERT)	0.09	0.12	0.08	0.13	Wilcoxon	.248	0.34

Glass-Box condition retained a smaller proportion of AI-generated tokens ($M = 82.45\%$) than in the Black-Box condition ($M = 84.58\%$). However, this difference did not reach the threshold for statistical significance ($p = .085$). Although the associated rank-biserial effect size was relatively large ($r = .52$), the absolute difference between conditions was modest, approximately two percentage points. Therefore, this result should be interpreted as a preliminary directional signal rather than evidence of a robust behavioural difference.

5.4. Analysis of prompt editing in the glass-box condition

To address the nature and extent of user interventions during the generative process, we analysed the manual modifications made to the LLM prompts in the Glass-Box condition. Behavioural logs revealed that 5 out of 30 participants (16.7%) actively edited the prompt before triggering the AI generation. Because these interventions were highly targeted and adopted by a specific subset of users, aggregate computational metrics (e.g. Levenshtein distance) would have been heavily skewed by the zero-values of the majority. Therefore, we conducted a qualitative thematic analysis to extract deeper insights into the users' intent. Two researchers independently analysed the edited prompts and resolved discrepancies through consensus, identifying three primary editing patterns.

The first pattern refers to the use of prompt editing to *increase contextual and operational specificity*. Rather than changing the overall scenario, these participants expanded underspecified prompts by clarifying exactly what the email should communicate and what action the recipient should take. For example, one participant transformed a brief baseline prompt for the 'Online Accounts' scenario into a more concrete instruction, explicitly stating that the AI should warn the victim about *'suspicious and unusual access'* and that the message *'requires some quick actions'*.

The second pattern concerns *surface-level output control*, especially regarding language and formatting. Users inserted explicit constraints that were not present in the baseline prompt. For instance, participant P12 added *'Scrivi la mail in inglese'* (Write the email in English) and *'Scrivi l'email in inglese e da parte di un dipendente del reparto IT…'* (Write the email in English as an IT department employee) in scenarios 3 and 4 to strictly enforce the output format.

The third pattern involves edits for *tone calibration and simplification*. In one case, a participant added *'Show yourself as worried for the user'* (P5) to make the intended emotional stance more explicit to the LLM. Conversely, other edits involved prompt simplification, where participants removed ancillary wording generated by the wizard while carefully preserving the core task definition.

6. Discussion

This study aimed to explore the design space of EUD for GenAI, specifically focussing on the trade-off between

transparency (Glass-Box) and *abstraction* (Black-Box). A common assumption in HCI is that exposing technical details increases cognitive load for non-experts (Zamfirescu-Pereira et al. 2023; Kulesza et al. 2013). Our results challenge this view, supporting a 'Transparency without Overhead' thesis (B. Y. Lim and Dey 2011; Kizilcec 2016). We discuss the implications below. Nevertheless, the findings should be interpreted in relation to the specific EUD mechanisms implemented in the study. The evaluated interface does not rely on explicit rule authoring or full programming by demonstration. Instead, it shows how EUD principles can be adapted to GenAI systems by treating the prompt as a scaffolded, editable configuration artifact.

6.1. RQ1: quality parity between paradigms

Our first question addressed whether the granular control offered by the Glass-Box approach might inadvertently lead users to degrade the output quality compared to a safer, constrained Black-Box approach. The analysis of the *Quality Index* indicates *statistical equivalence*. This finding suggests that the Glass-Box interface successfully supported users in managing the prompt. The scaffolding provided by the system prevented the 'blank page syndrome' often associated with raw prompting (Zamfirescu-Pereira et al. 2023), allowing users to achieve the same high-quality results (persuasion, realism, safety) as the opaque Black-Box wizard. In essence, exposing the mechanism did not break the machine (Wu, Terry, and Cai 2022).

6.2. RQ2: the myth of cognitive overload

We hypothesised that the Black-Box condition, by hiding complexity, would be perceived as more usable and less demanding. Surprisingly, the results demonstrate *cognitive parity*. The Glass-Box condition imposed no significant additional workload (NASA-TLX) and was rated equally usable (SUS). This supports the usability of the specific EUD-inspired mechanisms employed (scaffolded prompt configuration, real-time prompt visibility, and parameter-based selection) (Hutchins, Hollan, and Norman 1985). By making the prompt visible and immediately reactive, the Glass-Box interface may have helped users treat the prompt as a cognitive aid rather than as an additional burden, supporting their understanding of the link between selected settings and the AI's output.

Although the average SUS scores for both conditions were in the good range, their relatively large standard deviations suggest substantial inter-participant variability in perceived usability. Our observational notes indicate that lower usability ratings were often associated with initial difficulties in navigating the platform and configuring the phishing campaign, including uncertainty in selecting the correct scenario and in specifying emotional triggers and persuasion strategies. In a few cases, participants omitted a trigger, selected the wrong scenario, or manually edited the prompt to compensate for what they perceived as an inadequate default configuration. These observations suggest that the variance in SUS may reflect differences in onboarding and configuration ease rather than a uniform usability issue across the platform. However, since the study was not designed to test the effect of technical literacy or prior experience, we cannot conclude that a specific user profile systematically struggled with the system.

6.3. The role of transparency vs. active editing

It could be argued that the lack of significant differences in cognitive load (RQ2) and email quality (RQ1) stems from the fact that a large majority of participants (83.3%) chose to read, but not actively edit, the exposed prompt in the Glass-Box condition. However, this observation does not diminish the value of the EUD interface; rather, it highlights the efficacy of the *Scaffolded Transparency* approach. The fact that most users accepted the system-generated prompt indicates that the shared scaffolded configuration workflow (Steps 2a–2c) successfully translated high-level intents into usable LLM instructions, reducing the risk of 'blank page syndrome'. The cognitive parity observed in our results demonstrates that simply *exposing* the prompt – requiring users to read, process, and validate the AI's logic before proceeding – did not induce cognitive overload. Furthermore, while the prompt was rarely edited, the *AI Retention Rate* showed a non-significant trend in the expected direction. Participants in the Glass-Box condition retained slightly less AI-generated text than in the Black-Box condition. However, because this difference did not reach statistical significance and the absolute difference was modest, the result should be interpreted cautiously. Our data are compatible with the possibility that exposing the generative logic may be associated with a modest difference in textual reliance on AI-generated drafts, but this interpretation remains exploratory and warrants validation in future work.

6.4. RQ3: transparency, textual reliance, and user agency

The behavioural metrics provide a more nuanced picture than the quality and workload results. While

performance and workload were similar across conditions, the Glass-Box condition showed a non-significant trend toward lower *AI Retention Rate*. This pattern is compatible with the possibility that prompt visibility may have influenced how some users revised the AI-generated draft. However, because the difference was not statistically significant, AI Retention Rate should be interpreted only as a proxy for textual reliance, not as direct evidence of automation bias, changed mental models, or supervisory behaviour. This interpretation is theoretically consistent with prior work showing that reduced process visibility can foster over-reliance on automated decision aids (Bansal et al. 2021; Skitka, Mosier, and Burdik 2000). However, the present study does not directly measure over-reliance or automation bias. Therefore, we treat the observed AI Retention Rate pattern as an exploratory behavioural signal that may inform future studies using direct measures of verification behaviour, trust calibration, and mental-model change.

From a design perspective, the Glass-Box interface frames the AI less as an opaque generator and more as an inspectable drafting tool. This framing may support critical engagement with AI-generated content (Kizilcec 2016; Kulesza et al. 2013; Yuan et al. 2022); however, our study provides only indirect behavioural evidence for this possibility. Future work should directly measure whether prompt visibility changes users' verification strategies, trust calibration, or mental models of the system.

An alternative interpretation of the lower AI Retention Rate could be that the Glass-Box interface produced inferior initial drafts, thereby forcing users to edit the text out of necessity rather than active agency. However, the behavioural data do not provide clear support for this interpretation. As detailed in Section 5.4, 83.3% of participants did not manually alter the prompt in the Glass-Box condition. Consequently, the vast majority of the initial AI drafts in the Glass-Box condition were generated by the exact same underlying prompt logic as in the Black-Box condition, supporting the interpretation that baseline draft quality was comparable across conditions. Prompt transparency may have influenced how some users approached the AI-generated draft, but the present study does not directly measure the psychological mechanisms underlying this behaviour.

7. Implications for design and policy

The findings of this study extend beyond the specific domain of phishing simulations, offering broader actionable insights for the design of Human-Centered AI systems. We articulate three key implications for researchers, designers, and organisational policymakers:

(1) From 'Black-Box Simplification' to 'Scaffolded Transparency'. A prevailing heuristic in AI interface design assumes that to make complex models usable for non-experts, one must hide the underlying logic (the 'Black Box' approach) (Amershi et al. 2019; Liao, Gruen, and Miller 2020; Springer and Whittaker 2019). Our results challenge this dichotomy. We demonstrated that complexity does not necessarily equate to cognitive overload if the transparency is properly structured.

Design Implication: Developers of GenAI tools should prioritise *Scaffolded Transparency*. Instead of removing the prompt to simplify the interface, designers should expose it through visual scaffolding (e.g. live previews, modular prompt components, and parameter-based selection) (Arawjo et al. 2024; Tanimoto 2013; Wu, Terry, and Cai 2022). This approach bridges the 'Gulf of Evaluation' (Hutchins, Hollan, and Norman 1985), allowing users to build a correct mental model of the AI's behaviour without needing prompt engineering expertise. Transparency should be treated as a usability feature, not a technical hurdle.

(2) EUD as a Support for Reflective AI Interaction. The observed trend in the AI Retention Rate ($r = 0.52$) provides a preliminary indication that the interaction modality may influence how users preserve or revise AI-generated wording. While the absolute difference in retention rates between conditions was modest (82.45% vs. 84.58%) and did not reach statistical significance ($p = .085$), the associated effect size falls in the medium-to-large range (Cohen 2013). This pattern should be interpreted with caution: rather than indicating a strong or definitive behavioural shift, it constitutes a directionally consistent signal that warrants further investigation in longitudinal studies. The practical relevance of even small reductions in AI retention should not be dismissed outright, however, as in high-stakes domains such as security operations, marginal differences in critical engagement may compound meaningfully over repeated use.

Design Implication: To prevent over-reliance on AI in high-stakes decision-making, interfaces should avoid 'Magic Button' interactions that deliver instant, finished artifacts (Buçinca, Malaya, and Gajos 2021). Instead, designs should foster *Co-Creation Workflows* (Wu, Terry, and Cai 2022; Yuan et al. 2022) that encourage user engagement at the configuration stage. By transparently exposing the generative logic and allowing the user to inspect or manipulate the parameters *before* the output is finalised, the system creates a beneficial 'cognitive friction'. This visibility may introduce a form of

cognitive friction that invites users to inspect the generative process before accepting the output.

(3) Scaling Security Operations through 'Bounded Freedom'. For public administrations and large organisations, the centralisation of security training often leads to generic, ineffective campaigns (Ho et al. 2025). Conversely, full decentralisation risks non-compliance (Kirlappos, Parkin, and Sasse 2015). Our findings on safety non-inferiority provide a solution: EUD enables a model of *Bounded Freedom*.

Policy Implication: Organisations should adopt EUD-based platforms to democratise security operations. This approach enables local administrators, who possess the contextual knowledge (culture, language, and current events) necessary for realistic simulations, to tailor content freely. Meanwhile, the system enforces hard-coded safety constraints (e.g. prohibited payloads and mandatory educational disclaimers) in the backend. This 'Federalised' approach to AI governance maximises local relevance while guaranteeing global compliance.

8. Limitations

We acknowledge certain limitations in our study. First, while our sample size ($N = 30$) is consistent with typical HCI standards for usability studies, it limits the statistical power to detect minute differences. However, the *Null Results* regarding Quality and Usability are supported by *negligible effect sizes* (Cohen's $d<0.10$ for both Quality Index and SUS). This strongly suggests that the lack of statistical significance is not due to Type II errors (insufficient power) but reflects a *practical equivalence* between the two interfaces. In other words, even if a larger sample were to identify a statistically significant difference, the magnitude of that difference would likely be too small to have operational relevance in a real-world context.

Second, the study was conducted in a controlled laboratory setting. While this ensured internal validity and consistent model parameters, it may not fully capture the chaotic nature of real-world security operations. Future longitudinal studies could assess whether the 'Glass-Box' approach maintains its benefits over prolonged usage periods.

9. Conclusion

This study addressed a critical tension in the design of AI-assisted cybersecurity tools: the trade-off between ease of use through hidden prompt configuration (Black-Box) and user control through visible and editable prompt configuration (Glass-Box). Through the design and evaluation of DAMOCLES, we demonstrated that this trade-off is not inevitable. By leveraging EUD-inspired mechanisms such as scaffolded prompt configuration, real-time prompt visibility, and editable generative instructions, we achieved a condition of 'Democratisation without Degradation'. The Glass-Box interface empowered non-expert users to configure sophisticated phishing simulations with the same quality and safety as a rigid wizard, and with no additional cognitive cost.

Most importantly, our findings clarify the potential role of transparency in AI-assisted content generation. Exposing the prompt did not compromise quality, safety, or usability, and produced a preliminary, non-significant behavioural trend toward lower retention of AI-generated text. This result suggests a possible role for scaffolded transparency in supporting more reflective interaction with GenAI systems. However, because automation bias, mental models, and supervisory behaviour were not directly measured, future work is needed to verify these mechanisms explicitly. Future work will focus on longitudinal studies to assess whether scaffolded transparency affects users' reliance, verification strategies, and engagement over time, as well as on exploring the application of Glass-Box EUD to other defensive tasks, such as generating training scenarios or policy definitions.

Notes

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
5. *Safety* is defined as whether the generated phishing email violates ethical safeguards or global compliance.
6. Received on 27 November 2025, Prot. 0323653.

Author contributions

CRediT: **Giuseppe Desolda:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Writing – original draft, Writing – review & editing; **Vincenzo Deufemia:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Writing – original draft, Writing – review & editing; **Lucio Davide Spano:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Writing – original draft, Writing –

review & editing; **Paolo Buono:** Investigation, Methodology, Writing – review & editing; **Federico Maria Cau:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing; **Gaetano Cimino:** Data curation, Formal analysis, Investigation, Methodology, Writing – review & editing; **Matteo Mocci:** Formal analysis, Investigation, Methodology, Writing – review & editing; **Cesare Tucci:** Data curation, Formal analysis, Investigation, Methodology, Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research is partially funded by the Italian Ministry of Universities and Research (MUR) and by the, 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.

ORCID

Giuseppe Desolda http://orcid.org/0000-0001-9894-2116*Lucio Davide Spano* http://orcid.org/0000-0001-7106-0463

Declaration of generative AI

During the preparation of this work, the author(s) used *Grammarly Pro* in order to fix grammatical errors and improve text quality.

References

Abroshan, H., J. Devos, G. Poels, and E. Laermans. 2021. "Covid-19 and Phishing: Effects of Human Emotions, Behavior, and Demographics on the Success of Phishing Attempts during the Pandemic." *IEEE Access* 9:121916–121929. https://doi.org/10.1109/ACCESS.2021.3109091.

Agarwal, D., M. Naaman, and A. Vashistha. 2025. "AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances." In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25*. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3706598.3713564.

Akbarian, F., E. Fitzgerald, and M. Kihl. 2020. Split, Croatia.

Alexandrova, S., M. Cakmak, K. Hsiao, and L. Takayama. 2014. "Robotics: Science and Systems." Berkeley, California, USA.

Alhaddad, M., M. Mohd, F. Qamar, and M. Imam. 2023. "Study of Student Personality Trait on Spear-Phishing Susceptibility Behavior." *International Journal of Advanced Computer Science and Applications* 14 (5): 667–678. https://doi.org/10.14569/issn.2156-5570.

Alissa, K. A., H. A. Alshehri, S. A. Dahdouh, B. M. Alsubaie, A. M. Alghamdi, A. Alharby, and N. A. Almubairik. 2018. Saudi Computer Society National Computer Conference (NCC), Riyadh, Saudi Arabia.

Amershi, S., D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, et al. 2019. "Guidelines for Human-AI Interaction." In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19*, 1–13, New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3290605.3300233.

Arawjo, I., C. Swoopes, P. Vaithilingam, M. Wattenberg, and E. L. Glassman. 2024. "ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing." In *CHI Conference on Human Factors in Computing Systems*, New York, NY: Association for Computing Machinery.

Ardito, C., G. Desolda, R. Lanzilotti, A. Malizia, M. Matera, P. Buono, and A. Piccinno. 2020. "User-Defined Semantics for the Design of IoT Systems Enabling Smart Interactive Experiences." *Personal and Ubiquitous Computing* 24 (6): 781–796. https://doi.org/10.1007/s00779-020-01457-5.

Artizzu, V., G. Cherchi, D. Fara, V. Frau, R. Macis, L. Pitzalis, A. Tola, I. Blecic, and L. D. Spano. 2022. "Defining Configurable Virtual Reality Templates for End Users." *Proceedings of the ACM on Human-Computer Interaction* 6: 1–35. https://doi.org/10.1145/3534517.

Bansal, G., T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld. 2021. "Does the Whole Exceed Its Parts? The Effect of Ai Explanations on Complementary Team Performance." In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16. New York, NY: Association for Computing Machinery.

Breve, B., P. Buono, L. Caruccio, F. M. Cau, G. Cimino, G. Desolda, V. Deufemia, R. Lanzilotti, L. D. Spano, and C. Tucci. 2025. "Leveraging EUD and Generative AI for Ethical Phishing Campaigns." In *End-User Development*, edited by C. Santoro, A. Schmidt, M. Matera, and A. Bellucci, 264–282. Cham: Springer Nature Switzerland.

Breve, B., G. Cimino, and V. Deufemia. 2024. "Hybrid Prompt Learning for Generating Justifications of Security Risks in Automation Rules." *ACM Transactions on Intelligent Systems and Technology* 15 (5): 1–26. https://doi.org/10.1145/3675401.

Buçinca, Z., M. B. Malaya, and K. Z. Gajos. 2021. "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on Ai in AI-assisted Decision-Making." *Proceedings of the ACM on Human-Computer Interaction* 5: 1–21. https://doi.org/10.1145/3449287.

Buono, P., G. Desolda, F. Greco, and A. Piccinno. 2023. "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*, 1–6. New York, NY: Association for Computing Machinery.

Cimino, G., G. Carenini, and V. Deufemia. 2026. "MIMIC: Multi-party Dialogue Augmentation via Speaker Stylistic Transfer." In *Findings of the Association for Computational Linguistics: EACL 2026*, 2693–2719. Rabat, Morocco: Association for Computational Linguistics.

Cohen, J.. 2013. *Statistical Power Analysis for the Behavioral Sciences*. London: Routledge.

Corradini, I., and E. Nardelli. 2019. "Building Organizational Risk Culture in Cyber Security: The Role of Human Factors." In *Advances in Human Factors in Cybersecurity*, edited by T.Z. Ahram and D. Nicholson, 193–202. Cham: Springer International Publishing.

Desolda, G., J. Aneke, C. Ardito, R. Lanzilotti, and M. F. Costabile. 2023. "Explanations in Warning Dialogs to Help Users Defend against Phishing Attacks." *International Journal of Human-Computer Studies* 176:103056. https://doi.org/10.1016/j.ijhcs.2023.103056.

Desolda, G., L. S. Ferro, A. Marrella, T. Catarci, and M. F. Costabile. 2021. "Human Factors in Phishing Attacks: A Systematic Literature Review." *ACM Computing Surveys (CSUR)* 54 (8): 1–35. https://doi.org/10.1145/3469886.

Desolda, G., F. Greco, R. Lanzilotti, and C. Tucci. 2026. "Morpheus: A Multidimensional Framework for Modeling, Measuring, and Mitigating Human Factors in Cybersecurity." *ACM Transactions on Computer-Human Interaction*. https://doi.org/10.1145/3829371.

Eckhart, M., and A. Ekelhart. 2018. "Towards Security-Aware Virtual Environments for Digital Twins." In *Proceedings of the 4th ACM Workshop on Cyber-Physical System Security. CPSS '18*, 61–72. New York, NY: Association for Computing Machinery.

Faleiro, R., L. Pan, S. R. Pokhrel, and R. Doss. 2022. "Digital Twin for Cybersecurity: Towards Enhancing Cyber Resilience." In *Broadband Communications, Networks, and Systems*, edited by W. Xiang, F. Han, and T.K. Phan, 57–76. Cham: Springer International Publishing.

Ghiani, G., F. Paternò, L. D. Spano, and G. Pintori. 2016. "An Environment for End-User Development of Web Mashups." *International Journal of Human-Computer Studies* 87:38–64. https://doi.org/10.1016/j.ijhcs.2015.10.008.

Glaessgen, E., and D. Stargel. 2012. "The Digital Twin Paradigm for Future NASA and U.S. Air Force Vehicles." In *53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*. Honolulu: AIAA.

Greco, F., P. Buono, D. Desiato, G. Desolda, R. Lanzilotti, and G. Ragone. 2024. Arenzano: CEUR-WS.

Greco, F., G. Desolda, C. Tucci, A. Esposito, A. Curci, and A. Piccinno. 2025. "Improving Phishing Resilience with AI-Generated Training: Evidence on Prompting, Personalization, and Duration." https://arxiv.org/abs/2512.01893.

Guenduez, A. A., T. Mettler, and K. Schedler. 2020. "Technological Frames in Public Administration: What Do Public Managers Think of Big Data?" *Government Information Quarterly* 37 (1): 101406. https://doi.org/10.1016/j.giq.2019.101406.

Gupta, M., C. Akiri, K. Aryal, E. Parker, and L. Praharaj. 2023. "From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy." *IEEE Access* 11:80218–80245. https://doi.org/10.1109/ACCESS.2023.3300381.

Ho, G., A. Mirian, E. Luo, K. Tong, E. Lee, L. Liu, C. A. Longhurst, C. Dameff, S. Savage, and G. M. Voelker. 2025. "Understanding the Efficacy of Phishing Training in Practice." In *2025 IEEE Symposium on Security and Privacy (SP)*, 37–54. San Francisco, CA: IEEE. https://doi.org/10.1109/SP61157.2025.00076.

Hutchins, E. L., J. D. Hollan, and D. A. Norman. 1985. "Direct Manipulation Interfaces." *Human–Computer Interaction* 1 (4): 311–338. https://doi.org/10.1207/s15327051hci0104_2.

IBM Security. 2024. "IBM Security Data Breach Investigations Report 2024." Accessed November 15, 2024. https://www.ibm.com/reports/data-breach.

IFTTT. n.d. Accessed March 14, 2025. https://ifttt.com/.

Kim, B., D. Y. Lee, and B. Kim. 2020. "Deterrent Effects of Punishment and Training on Insider Security Threats: A Field Experiment on Phishing Attacks." *Behaviour & Information Technology* 39 (11): 1156–1175. https://doi.org/10.1080/0144929X.2019.1653992.

Kirlappos, I., S. Parkin, and M. A. Sasse. 2015. ""Shadow Security" as a Tool for the Learning Organization." *ACM SIGCAS Computers and Society*45 (1): 29–37. https://doi.org/10.1145/2738210.2738216.

Kizilcec, R. F. 2016. "How Much Information? Effects of Transparency on Trust in an Algorithmic Interface." In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*, 2390–2395. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2858036.2858402.

Koddebusch, M. 2022. "Exposing the Phish: The Effect of Persuasion Techniques in Phishing e-Mails." In *Proceedings of the 23rd Annual International Conference on Digital Government Research*, 78–87. New York, NY: Association for Computing Machinery.

Kulesza, T., S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W. K. Wong. 2013. "Too Much, Too Little, or Just Right? Ways Explanations Impact End Users' Mental Models." In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, 3–10. San Jose, CA: IEEE.

Kumaraguru, P., J. Cranshaw, A. Acquisti, L. Cranor, J. Hong, M. A. Blair, and T. Pham. 2009. "School of Phish: A Real-World Evaluation of Anti-Phishing Training." In *Proceedings of the 5th Symposium on Usable Privacy and Security*. New York, NY: Association for Computing Machinery.

Kumaraguru, P., S. Sheng, A. Acquisti, L. F. Cranor, and J. Hong. 2010. "Teaching Johnny Not to Fall for Phish." *ACM Transactions on Internet Technology* 10 (2): 1–31. https://doi.org/10.1145/1754393.1754396.

Li, T. J. J., I. Labutov, X. N. Li, X. Zhang, W. Shi, W. Ding, T. M. Mitchell, and B. A. Myers. 2018. "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*, 105–114. Lisbon: IEEE.

Liao, Q. V., D. Gruen, and S. Miller. 2020. "Questioning the AI: Informing Design Practices for Explainable AI User Experiences." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20*, 1–15. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3313831.3376590.

Lieberman, H., F. Paternò, and V. Wulf. 2006. *End User Development (Human-Computer Interaction Series)*. Berlin, Heidelberg: Springer-Verlag.

Lim, B. Y., and A. K. Dey. 2011. "Investigating Intelligibility for Uncertain Context-Aware Applications." In *UbiComp 2011: Ubiquitous Computing, 13th International Conference, UbiComp 2011, Beijing, China, September 17–21, 2011, Proceedings*, edited by J.A. Landay, Y. Shi, D.J. Patterson, Y. Rogers, and X. Xie, 415–424. Beijing: ACM. https://doi.org/10.1145/2030112.2030168.

Lim, I. K., Y. G. Park, and J. K. Lee. 2016. "Design of Security Training System for Individual Users." *Wireless Personal Communications* 90 (3): 1105–1120. https://doi.org/10.1007/s11277-016-3380-z.

Lyell, D., and E. Coiera. 2017. "Automation Bias and Verification Complexity: A Systematic Review." *Journal of the American Medical Informatics Association* 24 (2): 423–431. https://doi.org/10.1093/jamia/ocw105.

Mahdavi Goloujeh, A., Sullivan , A., and Magerko , B. 2024. New York, NY: Association for Computing Machinery.

Masson, D., Malacria , S., Casiez , G., and Vogel , D. 2024. New York, NY: Association for Computing Machinery.

Modic, D., R. Anderson, and J. Palomäki. 2018. "We Will Make You Like Our Research: The Development of a Susceptibility-to-Persuasion Scale." *PLoS One* 13 (3): e0194119. https://doi.org/10.1371/journal.pone.0194119.

Mosier, K. L., L. J. Skitka, M. D. Burdick, and S. T. Heers. 1996. "Automation Bias, Accountability, and Verification Behaviors." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 40 (4): 204–208. https://doi.org/10.1177/154193129604000413.

Mudassar Yamin, M., E. Hashmi, M. Ullah, and B. Katt. 2024. "Applications of LLMs for Generating Cyber Security Exercise Scenarios." *IEEE Access* 12:143806–143822. https://doi.org/10.1109/ACCESS.2024.3468914.

Myers, B., S. E. Hudson, and R. Pausch. 2000. "Past, Present, and Future of User Interface Software Tools." *ACM Transactions on Computer-Human Interaction* 7 (1): 3–28. https://doi.org/10.1145/344949.344959.

Ndibwile, J. D., Y. Kadobayashi, and D. Fall. 2017. "UnPhishMe: Phishing Attack Detection by Deceptive Login Simulation through an Android Mobile App." In *2017 12th Asia Joint Conference on Information Security (AsiaJCIS)*, 38–47. Seoul: IEEE.

Ovelgönne, M., T. Dumitraş, B. A. Prakash, V. S. Subrahmanian, and B. Wang. 2017. "Understanding the Relationship between Human Behavior and Susceptibility to Cyber Attacks: A Data-Driven Approach." *ACM Transactions on Intelligent Systems and Technology* 8: 1–25.

Parasuraman, R., and D. H. Manzey. 2010. "Complacency and Bias in Human Use of Automation: An Attentional Integration." *Human Factors: The Journal of the Human Factors and Ergonomics Society*52 (3): 381–410. https://doi.org/10.1177/0018720810376055.

Rizzoni, F., S. Magalini, A. Casaroli, P. Mari, M. Dixon, and L. Coventry. 2022. "Phishing Simulation Exercise in a Large Hospital: A Case Study." *Digital Health* 8:20552076221081716. https://doi.org/10.1177/20552076221081716.

Rossetti, G., M. Stella, R. Cazabet, K. Abramski, E. Cau, S. Citraro, A. Failla, R. Improta, V. Morini, and V. Pansanella. 2024. *Y Social: An LLM-powered Social Media Digital Twin*, arXiv preprint arXiv:2408.00818.

Sheng, S., B. Magnien, P. Kumaraguru, A. Acquisti, L. F. Cranor, J. Hong, and E. Nunge. 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*, 88–99. New York, NY: Association for Computing Machinery.

Skitka, L. J., K. Mosier, and M. D. Burdik. 2000. "Accountability and Automation Bias." *International Journal of Human-Computer Studies* 52 (4): 701–717. https://doi.org/10.1006/ijhc.1999.0349.

Spithoven, R., and A. Drenth. 2024. "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." *Journal of Cybersecurity* 10 (1): tyae010. https://doi.org/10.1093/cybsec/tyae010.

Springer, A., and S. Whittaker. 2019. "Progressive Disclosure: Empirically Motivated Approaches to Designing Effective Transparency." In *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI '19*, 107–120. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3301275.3302322.

Tanimoto, S. L. 2013. "A Perspective on the Evolution of Live Programming." In *Proceedings of the 1st International Workshop on Live Programming, LIVE '13*, 31–34. New York, NY: Association for Computing Machinery.

The Most Disruptive Trend of 2021: No Code/Low Code. n.d. Accessed March 14, 2025. https://www.forbes.com/sites/betsyatkins/2020/11/24/the-most-disruptive-trend-of-2021-no-code–low-code/.

Vaithilingam, P., et al. 2024. "DynaVis: Dynamically Synthesized UI Widgets for Visualization Editing." In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. New York, NY: Association for Computing Machinery.

Wen, Z. A., Z. Lin, R. Chen, and E. Andersen. 2019. "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*, 1–12. New York, NY: Association for Computing Machinery.

Workman, M. D., J. A. Luévanos, and B. Mai. 2022. "A Study of Cybersecurity Education Using a Present-Test-Practice-Assess Model." *IEEE Transactions on Education* 65 (1): 40–45. https://doi.org/10.1109/TE.2021.3086025.

Wu, T., M. Terry, and C. J. Cai. 2022. "AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts." In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–22. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3491102.3517582.

Xiong, A., R. W. Proctor, W. Yang, and N. Li. 2019. "Embedding Training within Warnings Improves Skills of Identifying Phishing Webpages." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 61 (4): 577–595. https://doi.org/10.1177/0018720818810942.

Yuan, A., A. Coenen, E. Reif, and D. Ippolito. 2022. "Wordcraft: Story Writing with Large Language Models." In *Proceedings of the 27th International Conference on Intelligent User Interfaces, IUI '22*, 841–852. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3490099.3511105.

Zamfirescu-Pereira, J. D., R. Y. Wong, Richmond, B. Hartmann, Q. Yang. 2023. *Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts*. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3544548.3581388.

Appendices

Appendix 1. Perceived control and transparency questionnaire

We measured participants' perceived control and transparency using the following two ad-hoc subscales:

- **Transparency**:
  1 I understood why the interface suggested certain formulations in the email text.
  2 It was clear to me which parts of the email were generated by the system and which depended on my choices.
  3 I could easily understand how the settings or parameters I selected influenced the resulting email text.
- **Control**:
  1 I felt able to guide the final outcome of the email.
  2 I could effectively intervene to modify the text to align with my intentions.
  3 I felt I had control over the psychological strategies used in the email.

Appendix 2. Selected scenarios

We thereby illustrate the four scenarios participants encountered during the study and assessed with both the **Glass box** and **Black box** interfaces. For each scenario, we explicitly asked participants to keep the message professional and believable so recipients do not immediately assume it is a test.

Scenario 1 – Banking/Financial
Persona and Context: You are Giuseppe Rossi, IT Manager at FinConsult S.p.A.

Goal: Write an email that appears to come from the bank's office, asking the recipient to perform an important account check shortly. The message must convey a *realistic concern* for potential problems if no intervention occurs (moderate and plausible fear), while communicating an authoritative and competent voice that simply explains what to do and why it is important to act.

NOTE: After the system has generated the draft based on the parameters you set, **reread and refine the email**. Ensure it achieves the scenario's goal and simply and precisely explains what the recipient must do (concrete steps and deadlines, if needed). Only after this revision select the users to send it to. This final check is a fundamental step of the study.

Scenario 2 – Online Accounts
Persona and Context: You are Maria Oxen, IT Manager at CampusX University.

Goal: Write an email informing about unusual access to the account and inviting the recipient to follow a quick security check. Convey a *controlled surprise* regarding the unexpected event and communicate that many colleagues have already done so, suggesting that following this procedure is the most natural and common choice. Maintain a reassuring and collaborative tone.

NOTE: After the system has generated the draft based on the parameters you set, **reread and refine the email**. Ensure it achieves the scenario's goal and simply and precisely explains what the recipient must do. Only after this revision select the users to send it to. This final check is a fundamental step of the study.

Scenario 3 – Online Shopping/Customer Service
Persona and Context: You are Giuseppe Rossi, IT Manager at RetailHub Italia.

Goal: Write a support email reporting a minor issue with an order and offering a small token of appreciation (such as a voucher or an upgrade) to reciprocate the recipient's cooperation in confirming a few non-sensitive details. The tone must convey positivity, encouraging a calm and helpful response.

NOTE: After the system has generated the draft based on the parameters you set, **reread and refine the email**. Ensure it achieves the scenario's goal and simply and precisely explains what the recipient must do. Only after this revision select the users to send it to. This final check is a fundamental step of the study.

Scenario 4 – Technical Support
Persona and Context: You are Giuseppe Rossi, CIO of Tech-Industries S.r.l.

Goal: Write an email that appears to be sent from the IT department, asking to perform a quick check to keep systems stable and functioning. The message must push the recipient to make a *personal commitment*, even using a firm and slightly annoyed tone regarding recent problems, while remaining respectful and professional.

NOTE: After the system has generated the draft based on the parameters you set, **reread and refine the email**. Ensure it achieves the scenario's goal and simply and precisely explains what the recipient must do. Only after this revision select the users to send it to. This final check is a fundamental step of the study.