RESEARCH-ARTICLE

Stop the Clock - Counteracting Bias Exploited by Attackers through an Interactive Augmented Reality Phishing Training

LORIN SCHÖNI, Swiss Federal Institute of Technology, Zurich, Zurich, ZH, Switzerland

MARTIN STROHMEIER, Armasuisse, Switzerland, Bern, BE, Switzerland

IVO SLUGANOVIC

VERENA ZIMMERMANN, Swiss Federal Institute of Technology, Zurich, Zurich, ZH, Switzerland

Published: 26 April 2025

Citation in BibTeX format

CHI 2025: CHI Conference on Human Factors in Computing Systems
April 26 - May 1, 2025
Yokohama, Japan

Conference Sponsors:
SIGCHI

Open Access Support provided by:

Armasuisse, Switzerland

Swiss Federal Institute of Technology, Zurich

CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (April 2025)
https://doi.org/10.1145/3706598.3714023
ISBN: 9798400713941

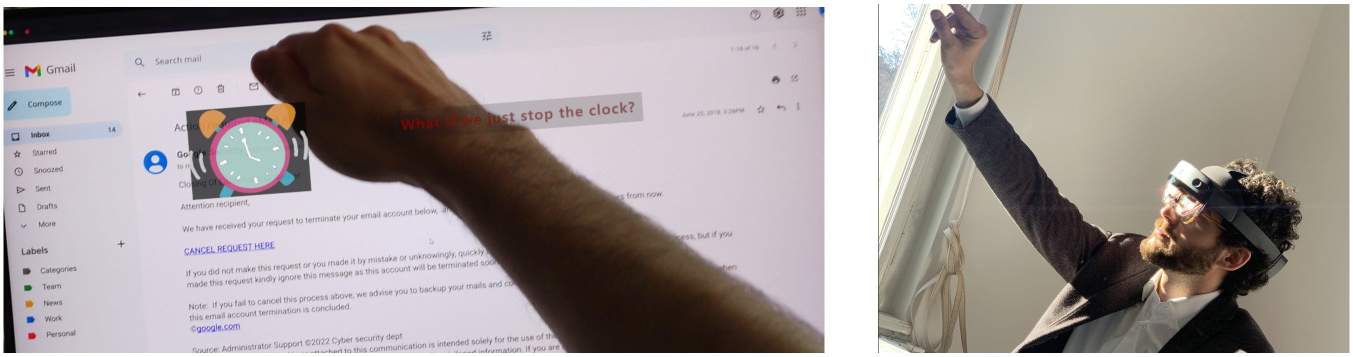
Stop the Clock - Counteracting Bias Exploited by Attackers through an Interactive Augmented Reality Phishing Training

Lorin Schöni
Security, Privacy & Society
ETH Zurich
Zurich, Switzerland
lorin.schoeni@gess.ethz.ch

Martin Strohmeier
Cyber-Defence Campus
armasuisse
Thun, Switzerland
martin.strohmeier@armasuisse.ch

Ivo Sluganovic
PhishAR
Oxford, United Kingdom
ivo.sluganovic@gmail.com

Verena Zimmermann
Department of Humanities, Social and Political Sciences
ETH Zürich
Zürich, Switzerland
verena.zimmermann@gess.ethz.ch



(a) View of the AR training experience captured through the AR headset.

(b) View of the user.

Figure 1: When looking at their computer screen, a user's view of the world is augmented with precisely positioned holographic visualizations of the biases, heuristics and norms most commonly exploited by attackers. Users are instructed to counter these biases by interactively responding to visual cues using natural hand interactions: e.g., by touching a holographic clock shown in space to counter the *urgency* that a phishing email is trying to induce and exploit.

Abstract

Phishing attacks become increasingly sophisticated in targeting humans and exploiting cognitive biases, e.g., through inducing authority or urgency. Previous approaches to user training focused on URL warnings, textual, or click-based training, yielding mixed results. For more interactive training, uncoupled from users' screens, we explore the potential of Augmented Reality (AR) technologies to enhance phishing detection. Through visual representations of biases that attackers typically exploit and gesture-based interactions with them, the training aims to enable users to counteract cognitive biases by increasing awareness and suspicion. In a laboratory study with $N = 117$ users, we evaluated phishing detection rates, user interaction with, and feedback on the AR-based training in comparison with a click-based variant and a control condition. Our results show that interactive phishing training addressing cognitive biases increased detection rates by 33% and that interactive elements were well perceived. AR technologies further enhance the training.

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3714023

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Security and privacy** → *Usability in security and privacy*; **Phishing**.

Keywords

Phishing, Augmented Reality, Training, Human-Centred Security

ACM Reference Format:
Lorin Schöni, Martin Strohmeier, Ivo Sluganovic, and Verena Zimmermann. 2025. Stop the Clock - Counteracting Bias Exploited by Attackers through an Interactive Augmented Reality Phishing Training. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3706598.3714023

1 Introduction

Phishing is highly prevalent with rates exceeding a million cases in the third quarter of 2022 alone [4]. Due to the increasing availability of sophisticated phishing kits that streamline attacks [8, 55], these threats are growing in scale and damage [40, 71, 94, 98]. Phishing attacks rely on deception, presenting malicious content as legitimate to trick human targets into taking dangerous actions [79], such as downloading an infected attachment, inputting credentials, or visiting a malicious website.

Technical phishing countermeasures aim to prevent users from interacting with threats [23, 101]. While useful, these solutions are in a constant arms race with attackers and cannot reliably prevent all threats on their own [45]. For example, approaches such as blocklists are insufficient due to the highly time-sensitive and dynamic nature of attacks [75] and an unfavourable cost-benefit ratio (e.g., [37]). Furthermore, phishing attacks often directly target humans, e.g., through social engineering and the exploitation of human cognitive biases and heuristics [11]. In the keynote for CHI 2023, Wolfangel prominently described the crucial role of the human in countering these threats [98].

Cognitive Biases Exploited by Attackers. Cognitive biases are heuristics and norms that are usually helpful mental shortcuts for decision-making in daily life. However, phishing attacks abuse these shortcuts by triggering quick and unsystematic thinking, tricking users into pattern-based, yet undesired and dangerous actions [41, 48]. For instance, an email could invoke urgency or present itself with an authority, such as a supervisor or a government agency, to manipulate users into unquestioningly following the request or emails. Even if a person knew of phishing and how to detect it, they might never even engage the sort of systematic thinking required to identify an email as phishing.

Human-Centred & Interactive Approaches to Phishing. To counteract the phishers' attack strategies and to complement technical solutions that are insufficient on their own, HCI researchers suggest more human-centred design approaches that support security-enhancing behaviour by considering human aspects such as users' perceptions, behaviors, or information processing [69, 98, 105]. In contrast, previous interventions often focused on compliance, used deception, or enforcement without considering human behaviour and cognition [92, 102]. Accordingly, many of these approaches lacked long-term effectiveness, suffered from low acceptance and knowledge gain, and were associated with high implementation costs [13, 31, 92].

In general, passive, text-based educational material has been shown to be inferior to interactive and multimedia material [13, 64]. Thus, our approach aims to increase active interaction with the training content to effectively enhance awareness for the cognitive biases, heuristics and norms exploited by attackers. The lacking interactivity and engagement could be addressed with emerging interactive technologies [45] such as Augmented Reality (AR), that allow for immersive, hands-on learning experiences.

Towards an Augmented Reality Anti-Phishing Intervention. Extended Reality technologies such as AR have demonstrated high potential to increase awareness [3, 72, 99], which is a crucial requirement for successful phishing training [64, 70]. Instead of exclusively focusing on either human or technological factors, AR enables their interaction and thus has the potential to bridge the gap between humans and technology for more effective training [6, 46] and more enduring effects (e.g., [64]). For example, AR-based training can increase interaction depth by making use of gestural interactions and visualizations that are superimposed over physical objects or screens in a second layer. Yet, even though an initial AR approach towards phishing education has been proposed [15], actual use of AR in phishing training is scarcely researched. Therefore, our research explores the potential of AR-based interactive phishing training that focuses on the cognitive biases exploited by attackers, to increase awareness for and, thereby, also detection rates of potential phishing emails. Crucially, such an AR training could flexibly be overlaid as a second layer over any other device without modifying work tasks while offering "in-the-moment" learning directly based on a user's work context.

Research Questions. This research was guided by the following research questions (RQs):

- **RQ1:** What are the effects of a training that directly targets cognitive biases and user attention towards suspicious cues in phishing emails in terms of a) users' awareness for the cognitive biases, heuristics and norms exploited by attackers, b) phishing detection ability, and c) users' evaluation of the training?
- **RQ2:** To what extent can the use of interactive AR-based training enhance the effects of the phishing training as compared to a) traditional text-based training or b) interactive but not-AR-based training?

To answer these research questions, we designed three variants of phishing trainings focusing on increasing user awareness for the cognitive biases, heuristics and norms exploited by attackers: a) an interactive AR-based training, b) a non-AR click-based training similar to [103] who compared an AR-based intervention with a 2D variant, and c) an additional text-based training as a control condition. To inform the training design and content that was similar across the three conditions, we first conducted a workshop with $N = 6$ cybersecurity experts. Based on the insights, we iteratively developed and implemented a training approach in which relevant biases, heuristics, and norms exploited by attackers were represented through visual and auditory cues in potential phishing emails. In the interactive AR-based training, users interacted with these cues using gestures to counteract them, e.g., by hitting an augmented alarm clock representing urgency (see Figure 1a). They were then informed about the reasoning of each interaction (see Fig. 2). In contrast, in the non-AR click-based training, users interacted with the same cues with click-based interactions, whereas, in the control condition, the same content was delivered in a textual form.

In a laboratory study with $N = 117$ users, we compared the effects of the three training conditions in a between-subject deign. We found that all training conditions substantially improved users' phishing detection rates, with the AR-based and click-based interactive conditions showing larger effects than the non-interactive textual control condition. Furthermore, the results showed that the AR-based condition leads to improved cybersecurity awareness and engagement compared to the other conditions, demonstrating the multi-faceted benefits of using the rich learning experience offered by an AR phishing training to enhance secondary outcomes.

The main contributions of this paper are:

- The research shows that phishers' exploitation of human cognitive biases, heuristics, and norms can be effectively countered by training enhancing users' awareness for and interaction with cues triggering those cognitive biases in phishing emails. The results suggest that users' awareness can effectively raise suspicion and trigger systematic thinking that enables users to better detect phishing.
- We proposed and implemented a novel and innovative approach that leverages AR technology to enhance engagement with phishing training, combining interactive technology and human-centred aspects to address the persistent challenge posed by phishing attacks. We make the source code of the phishing training's web application and AR application available for other researchers and practitioners to explore and adapt to their needs.
- The user study, by systematically comparing different degrees of interactivity, confirms the benefits of interactive training approaches and highlight the potential of AR as an emerging technology to further enhance interactivity. While AR-based training benefits still varied based on the users' technical affinity, the findings illustrated the AR-based intervention's potential to improve phishing detection while enhancing cybersecurity awareness and user engagement.

2 Related Work

We first review insights from human-centred anti-phishing interventions with regards to the role of training interactivity, suspicion and training focusing on biases, heuristics, and norms. We thereby outline how our AR-based training approach considers and extends these. Afterwards, we describe the relevant work related to the use of AR in cybersecurity.

2.1 Anti-Phishing Interventions

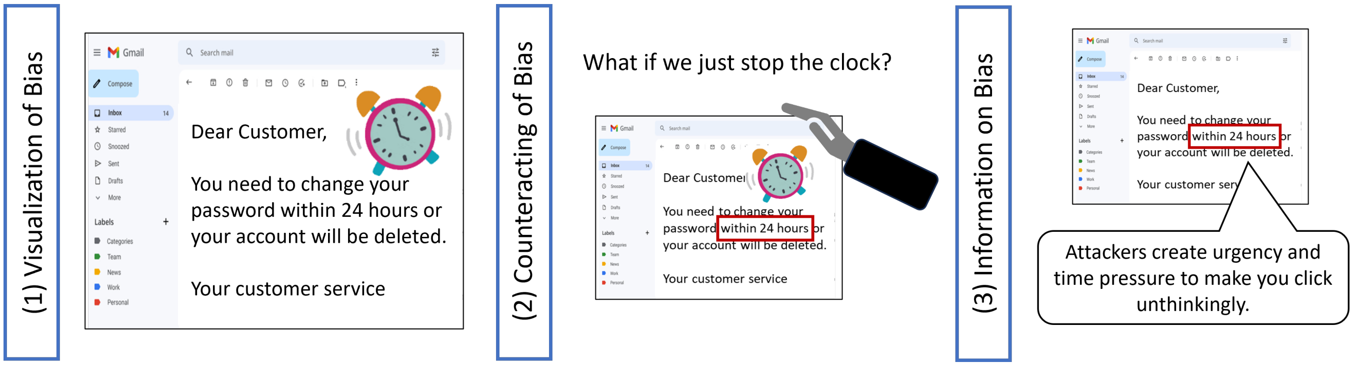
The role of training interactivity for training effectiveness. Research on phishing education by Wash et al. [90] demonstrated that advice-like education material can reduce click rates in phishing emails by 21%. However, users may lack the motivation to engage with non-interactive educational material in the first place [13]. Similarly, Sheng et al. [73] found that only users of interactive as compared to non-interactive training improved in a phishing classification task. Overall, interactive or embedded interventions were shown to be more effective than passive educational material [45, 73]. Increased interactivity, e.g., through gamification elements or serious games, effectively enhances phishing detection [74, 82, 92]. Therefore, our training approach includes interactive elements, but will be compared to a non-interactive variant to evaluate the assumed benefit.

The role of suspicion for phishing detection. Lin et al. [47] investigated domain highlighting techniques and found that most participants did not process incorrect domains even if they were specifically highlighted to draw visual attention. Further research confirmed these findings, with domain highlighting proving ineffective for phishing prevention [53, 66, 85, 100]. However, when users' attention is nudged [85] or forced [53] towards the URL, their phishing detection appears to improve. This indicates that users rarely look at details such as domains to check the authenticity of a website. However, once they have been given a reason to check the URL, phishing detection seems to improve. Wash [89] found that IT experts detect phishing emails in a similar way. The results showed that experts only become suspicious when noticing phishing cues, leading to a mindset that was conducive to detect phishing. Crucially, most experts did not investigate conclusive indicators like an URL until a sufficient number of cues that "seemed off" were identified. Our training thus similarly aims to raise suspicion that allows for a mindset of systematic processing, e.g., relevant for analysing URLs, through increased awareness for phishing cues in emails.

Vishwanath et al. [84] demonstrated that attention towards urgency cues was more likely to lead to increased elaboration on a phishing email. In contrast, attention to email source or grammar was much less likely to trigger such a response. Building on these findings, Vishwanath et al. [83] later proposed and evaluated the Suspicion, Cognition, and Automaticity Model of Phishing Susceptibility (SCAM). The model assumes suspicion to be the sole direct predictor of phishing susceptibility, which in turn is influenced by heuristic versus systematic information processing and email habits.

Training suspicion to counteract cognitive biases. The exploitation of human cognitive biases, heuristics, and norms is particularly problematic because it is inherent to our thinking. Hence, field experiments persistently demonstrate high phishing susceptibility [41] at around 20%, irrespective of context [76]. Yet, this effect can be counteracted by increasing awareness for specifically cognitive biases, which in turn increases suspicion towards phishing emails, enhances systematic information processing, and ultimately aids in phishing detection [11, 83, 89]. Therefore, our research aims to explore novel ways for enhancing users' awareness for the cognitive biases, heuristics, and norms that are exploited by attackers in phishing emails. This awareness is a relevant step towards increasing phishing detection rates by invoking suspicion and triggering systematic thinking, which in turn allows users to engage in learned detection strategies, such as analysing an URL.

The lack of training focusing on biases, heuristics, and norms. Only a few projects aim to capitalise on these factors. For example, Hashmi et al. [35], conducted a study to teach participants how to recognize persuasion principles to reduce their voice phishing susceptibility. In the training, 21 students listened to five voice recordings where a person was targeted in simulated phishing calls, using authority cues to trick their targets. However, the authors did not specifically educate participants about cues. When compared to a traditional awareness-raising method, the authors were unable to demonstrate significant benefits. Our study differs from this research, as we focus on interactive engagement with the training content, and provide targeted information on various cognitive biases represented through visual and auditory cues. Furthermore, our approach leverages the potential of AR that is detailed in the following.


(1) Visualization of Bias
Gmail
Search mail
Compose
Inbox 14
Starred
Snoozed
Sent
Drafts
More
Labels
Categories
Team
News
Work
Personal
(2) Counteracting of Bias
What if we just stop the clock?
(3) Information on Bias
Attackers create urgency and time pressure to make you click unthinkingly.

Dear Customer,

You need to change your password within 24 hours or your account will be deleted.

Your customer service

Figure 2: We investigate the effectiveness of a human-centred and interactive phishing training using AR technology. By visualizing the targeted biases, such as urgency visualized through a ticking clock (1) and having users actively counteracting them, the training aims to make strengthen users' processing. By actively stopping the ticking clock by hitting it (2), users counteract the bias in a metaphorical way and learn about the reasoning for the intervention (3).

2.2 Augmented Reality in Cybersecurity Research & Training

The potential of AR. AR training is widely used in the medical and health [6, 103], manufacturing [88], and education [14] domains. While effects are mixed, AR has been used to provide flexible and visual training, increase understanding, explain abstract concepts, and provide more interactive and captivating content [15, 17, 54, 59, 67]. AR also enables unique opportunities for more dynamic and immersive training [68, 93], effectively enhancing user engagement (e.g., [44, 68, 91]) and motivation [103]. For example, Kaiser et al. [42] used a mobile AR solution to assist users in decision-making through privacy visualisations in an emulated shopping scenario.

As AR can bridge between human and technical aspects, it offers a wide range of potential applications. As a second layer that is independent of other devices it can flexibly superimpose relevant digital information tailored to different concepts. The nature of wearable AR devices enables a wide range of interactions (e.g., gesture-based or gaze-based) and enhanced real-time immersive feedback [27, 59, 62]. These characteristics are ideal for interactive training, which can function best through user engagement (e.g., [44, 68, 91]).

Use of AR in cybersecurity training. While AR has been proposed for phishing education [15], AR applications that directly target cybersecurity and enhance human interaction to combat such threats remain sparsely researched. AR can be particularly effective in awareness training [3, 99], where traditional approaches show limited success (e.g., [46]). AR can also support the transfer of training-specific expertise to real life circumstances [6]. Therefore, calls have been made for more immersive and engaging technologies like AR to increase the effectiveness of such training [1, 93].

AR-based training appears most effective when information transfer is heavily human-centred and directly targets cognitive processes [6], such as with phishing detection. Recent research with VR environments have demonstrated that immersive interface design can both enhance user focus and introduce ergonomic challenges during phishing detection tasks [72]. AR applications in other areas (see [68]) have shown that they can decrease cognitive load, enhance focusing on important aspects, and are also more emotionally engaging, leading to higher user acceptance [17, 67]. We therefore expect the more involved nature of the AR training to especially increase engagement with the topic [30], which would thereby also increase people's inherent motivation to internalise the learned content and lead to higher training effectiveness.

2.3 Summary

Overall, the related work highlights a) the relevance of triggering suspicion to interrupt heuristic processing and to decrease phishing susceptibility, and b) the increased and longer-lasting effectiveness of interactive elements as compared to passive education material [11, 64, 93]. Therefore, we extend existing phishing research by evaluating a novel AR-based phishing training that targets cognitive biases in an interactive and engaging way to increase suspicion and trigger systematic thinking. Systematic thinking enables users to engage in further analysis of the email, thereby enhancing phishing detection. Our approach can be seen as complementary to techniques that focus on domain analysis, since triggering suspicion represents a first step towards focusing attention and motivating users to engage in further analysis [83].

3 Development of an Augmented Reality Phishing Training

To develop a human-centred phishing training using AR technology, we followed a process consisting of 1) an expert workshop for creating the training content (see Section 3.1), 2) an iterative implementation and evaluation phase (see Section 3.2), and 3) a large-scale user study (see Section 4).

3.1 Expert Workshop

In an expert workshop, N=6 experts for human-centered cybersecurity compiled and discussed a list of cognitive biases typically exploited by attackers in phishing emails. Of the N=6 experts, five had a background in psychology and one in criminology. Furthermore, N=3 held in-depth expertise in phishing. Three of the experts had completed a doctorate, one was pursuing a doctorate, and two were research assistants in the area of human-centred cybersecurity. The h-index of the four experts that had published papers in that domain ranged from 2 to 17 covering a range of 5 to 67 publications. To select suitable biases and to gather ideas on how to represent and interact with them in a training, the experts brainstormed, evaluated, and competitively compared different options based on insights from their own and related work. As such, there are works that investigated the psychological aspects or persuasion principles misused by attackers in the form of cues in phishing emails to evoke an automated response (e.g., [25, 25, 41, 95–97]. Specifically, the experts' collection of these principles and cues was closely based on and thus aligned with the works by Ferreira et al. [20, 21] who studied in which form persuasion elements are used in phishing and Gragg [32] who studied the psychological triggers behind social engineering to counteract phishing.

Procedure. The expert workshop was structured in four distinct phases, lasting a total of two hours. In a first phase, experts identified key biases that are commonly and effectively abused in phishing attacks in a collaborative, consent-oriented process. These biases were identified from both existing literature, the industry, and personal experience. If some overlap was identified, such as with urgency and scarcity, the biases were clustered together. In the second phase, each expert separately listed words or symbols that could be used to represent the cues in a training. For example, urgency creates time pressure and prompts rapid, instinctive reactions, represented by a ticking clock or a running person. Afterwards, all experts rated these options to identify the most suitable and understandable ones. In the third phase, experts then noted down interactions intended to help people understand and mentally counteract each option, e.g., hitting a ticking clock to stop it. To not restrain ideas, the experts were not considering any technical limitations that might hinder actual implementation. In the final phase, the experts then reflected on their choices in a discussion.

Findings. The key findings are summarized in Table 1. The psychological cues invoking biases were grouped into four meta categories: affect/emotion, needs and rewards, social influence, and context. We ensured that the identified psychological cues and persuasion principles were also supported by the literature: For example, Gragg [32] also identified strong affect and emotions such as fear or excitement as a relevant distraction that prevents the user from systematic information processing. Likewise, overloading, i.e., having to process lots of information quickly, affects logical reasoning and can lead users to become mentally passive [32]. Based on five persuasion principles, Ferreira et al. [20, 21] also identified the use of wording inducing fear, urgency, or authority as implementations of the persuasion principles "Authority" and "Distraction" (urgency also as an implementation of "Commitment, Reciprocation, and Consistency"). Providing an authentic and plausible context, e.g., through known graphics and logos, is identified as a relevant implementation of the principles "Authority", "Liking, Similarity & Deception" and "Distraction" [21]. Social cues, such a installing trust, providing social proof, reciprocity or helpfulness, are implementations of the principles "Liking, Similarity & Deception" and "Social Proof" [21], or "Reciprocation" as labeled by Gragg [32]. Additionally, Zielinska et al. [104] and Akbar [2] identified scarcity as a relevant cue in phishing emails analyzing phishing data sets.

Out of these categories, we further selected biases through an iterative process of considering technical feasibility and consulting with experts. Some biases, particularly in the social influence category, were considered relevant but infeasible for the intended laboratory study design. For instance, social biases are highly context- and person-specific and would be challenging to implement without individualised spear-phishing emails. After considering a combination of such factors, we finally decided on 4 biases and 3 representations for each of them. They served as a starting point for developing an AR-based and interactive training targeting cognitive biases and for testing whether increasing awareness of biases through an interactive approach can reduce phishing susceptibility and enhance secondary outcomes.

3.2 Iterative Development & Technical Implementation

In this section, we provide information about the prototype that was built to investigate and evaluate the effectiveness of interactive phishing training using AR technologies.

Existing AR applications mostly focus on representing or emulating concrete real-world objects, whereas abstract concepts are less often visualized. In particular, our application aimed to superimpose virtual visual elements (AR holograms) over the specific real-world object (user's monitor) and interact with the email content shown on it, which required keeping their location static and stable.

In order to compare the AR intervention to a click-based intervention, we required a non-AR application that uses the same elements, but displays them directly on the user's monitor. As such, we developed two separate applications: A web-based application that can easily be scaled to any number of users, and an AR-based application that then interfaces with the web-based application to show the interactions as virtual holograms.

The prototype thus consists of three main components: **a)** the web-based editor that allows researchers to define scenarios described in Section 4.3, **b)** the web application that displayed scenarios to users during experiments, and **c)** the AR application that presented visualizations and enabled interactivity. The web-based applications (a) and (b) were built using Vue [87] as the front-end that communicates with a node.js [22] back-end server, which also relays events and instructions to the AR application. It can be used to load screenshots and other elements like annotations, images, and sound. Through a hierarchy of conditions, navigation between the elements and slides was established. The prototype of the proposed AR application was built using the Unity Game Engine [81] and Mixed Reality Toolkit 2 [52]. It was deployed on Microsoft HoloLens 2 [51], which is capable of precisely overlaying a user's view of the physical world with virtual holograms, and supports natural hand interactions due to its advanced hand tracking capabilities.

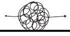
Source Code. In order to enable others to expand upon the prototype system built for this research, we have published the full code of the system implementation.[1]

[1] GitHub Repository: https://github.com/lorinschoeni/augmented-phishing

Table 1: Most Relevant Cues used to activate Cognitive Biases and Exemplary Representations and Interactions Identified in the Expert Workshop. The Cognitive Biases Selected for the User Study Are Marked with an Asterisk *.

Category	Bias	Example Representation	Exemplary Interaction
Affect/Emotion, e.g., [32]	*Fear	Ghost	Hit the ghost
	*Happiness	Cheerful sun	Turn the light off
	Curiosity	Box	Shake the box, clown jumps out
Social Influence, e.g., [20, 21, 32]	Attachment & Unity	Network	Cut into pieces
	Reciprocity & Helpfulness	Helping hand	Pull to see what hides behind
	Commitment & Consistency	Contract	Magnifying details of contract
	Authority	Large and small person	Making the large person smaller
	Trust & Social Proof	Picture of family	Reveal evil faces
Context, e.g., [20, 21, 32]	*Overloading/Confusion	Smoking head	Splashing with watering can
	Authenticity & Plausibility	Puzzle with hole	Seeing the piece does not fit
Needs & Rewards, e.g., [20, 21, 32, 104]	*Urgency	Ticking clock	Hit to make it stop
	Scarcity	Almost-empty shelf	Looking behind shelf
	Financial Incentives	Money	Vanishes once touched
	Need Fulfillment	Maslow's Pyramid	Identify correct need

Table 2: Training Email Content and Cognitive Biases

E-Mail Content	Cognitive Bias	Illustration	Sound	Interaction
Update Social Security Information	Fear		boo sound	Touch the ghost to make it disappear
Account Deletion Warning	Urgency		ringing sound	Hit the alarm clock to stop it from ringing
Alumni Party with Free Drinks	Happiness		party music	Reveal the happy smiley as lying and stop the music with a record scratch
Exciting Opportunity	Confusion		-	Use a watering can to calm the head down
Facebook Copyright Violation	Urgency		running sound	Stop the figure from running
Savings on Electricity Bill	Happiness		upbeat music	Reveal the sun as a fake paper sun
Emergency Contact for Injured Person	Fear		thunder sound	Make the clouds disappear to reveal a clear sky
Reception of Gift Card	Happiness		fireworks	Touch the balloon to make it pop and stop the fireworks
Fraudulent Paypal Transactions	Fear		spider sounds	Hit the spider
Government Research Survey	Confusion		-	Find the right key to open the chest
Free Class for Limited Time	Urgency		ticking sounds	Stop the clock
Human Resources Information	Confusion		-	Untangle the string

4 User Study

To explore the effects of the interactive AR-based training intervention as compared to a non-AR click-based training and a text-based control condition with the same training content, we conducted a laboratory between-subject study with N=117 people. Participants were randomly assigned to one of the three training interventions. After one and three weeks each, we explored long-term effects of the training in follow-up surveys. As dependent variables, we measured (a) the effect on phishing detection, (b) the user interaction with the training intervention, and (c) the user evaluation of the training intervention. The following sections detail the study procedure and material, its technical setup and methods of data collection, and finally related ethical considerations and the sample.

4.1 Procedure

The research comprised an initial laboratory study with a training module, followed by two remote online surveys. The procedure is summarised in Fig. 3. **(1)** After participants provided informed consent, they filled out the pre-intervention questionnaire, which included demographics, the Security Behaviour Intention Scale (SeBIS, [18]), the Affinity for Technology Interaction Scale (ATI, [24]), questions on experience with phishing and phishing training, and on the participant's mental model of phishing. Afterwards, we calibrated the eye tracking that was used for exploring users' attention and cognitive load related to the training and study content. Then, participants conducted a phishing classification task (see Section 4.3 for more detail). **(2)** Afterwards, they proceeded to one of the three training conditions they have been assigned to (see Section 4.2). **(3)** After navigating through all 12 sections, participants again conducted an email classification task, so we could compare behaviour before and after the intervention. Afterwards, we measured participants' mental task load with the NASA Task Load Index (NASA-TLX, [34]). We furthermore collected training feedback via the System Usability Scale (SUS, [10]), the short form of the User Engagement Scale (UES-SF, [58]), and self-created questions asking for interactivity, enjoyment, and user experience. Finally, we asked for users' mental model of phishing as well as their self-reported increase in awareness to explore the interventions' impact on these aspects. The complete set of questionnaires can be found in Appendix A.

(4) & (5) The laboratory study was supplemented by two follow-up online studies. Each participant was sent a link to start the study at the same time, which they could access within 48 hours, one week and three weeks after the laboratory study took place. At the start of both online studies, participants were again shown an informed consent sheet. Afterwards, a shortened questionnaire was used, including a phishing classification task and user evaluation in both follow-ups (see Appendix A).

4.2 Training

Before the training, all participants received a short introduction on phishing and cognitive biases (see Appendix B). Afterwards, they were assigned to one of three training variants, illustrated in Fig. 4. Throughout the training, participants interacted with 12 sections, each showing a phishing email. In the interactive click-based and AR-based conditions, they included elements supposed to make the cognitive biases heuristics triggered by attackers more graspable, by representing them through sound, images, gifs, boxes, and text elements. Participants interacted with these elements to metaphorically counteract the biases. These media elements then appeared or disappeared based on user interaction and were finally replaced by a short information on the reasoning for the intervention. In contrast, in the control condition, the screenshots were only accompanied by textual information on the cues. Afterwards, users proceeded to the next section until all 12 were completed. See a) in Fig. 4 for an example. Notably, all conditions mirrored each other in content. Only the presentation, medium, and degree of interactivity was manipulated. An overview of interactive training sections is given in Table 2.

4.3 Email Classification Task

Before and after the training intervention, participants conducted a phishing classification task with 25 fictional email examples that were randomly drawn from a larger set of 50 emails. This allowed us to randomise the sequence of emails, reducing the impact of repetition and potential differences in the difficulty of the classification task. Of the 25 emails shown, 10 were always designed as phishing emails with respective cues such as fake email domains or nonsense attachments.

These emails were either adapted from existing studies [12, 56, 63] or modeled after real-world occurrences and adapted for the study context. To enhance the contextual relevance of emails, we adapted their content to match participants' local context, such as adjusting names of senders and locations, or imitating local brands. Finally, a small number of phishing emails were manually created. We ensured that both phishing and non-phishing emails contained a comparable number of cues abusing cognitive biases, to control for potential over-suspicion that leads to participants classifying all emails as phishing if they appear to abuse cognitive biases. Participants classified these phishing emails prior to the training into either phishing or non-phishing. We provide detailed results and an overview of cognitive biases in Table 6 in Appendix C. The regular emails were classified as phishing with a rate of 0.33 (SD = .21), while phishing emails were classified as phishing with a rate of 0.66 (SD = .17). The rates indicate that participants were able to clearly separate between the two, yet, also show a high task difficulty level with variability between emails.

The follow-up surveys repeated this task but contained 24 emails each of which 16 where phishing. Out of these, eight were variations of training emails including the cognitive biases targeted on the training and eight were variations of the previous classification task emails. Four of the eight non-phishing emails also contained cognitive bias cues.

In the laboratory study, we chose to show more legitimate emails than phishing emails to create a perception that the majority of emails were legitimate and to avoid a predictable split into equally sized groups. In the follow-up surveys, we used a different ratio to minimise participant effort while including multiple examples for each cognitive bias represented in the training.

To emulate realistic circumstances and to induce the quick heuristic thinking processes that attackers abuse, and which the training aims to interrupt by raising awareness for the cognitive biases, participants were prompted to classify emails "as fast and precise as possible". After the classification task, participants answered questions on cognitive effort of the task using the NASA Task Load Index (NASA-TLX, [34]).

4.4 Technical Setup

The training was facilitated through a web app for both the click-based and AR-based conditions, and an additional HoloLens 2 app

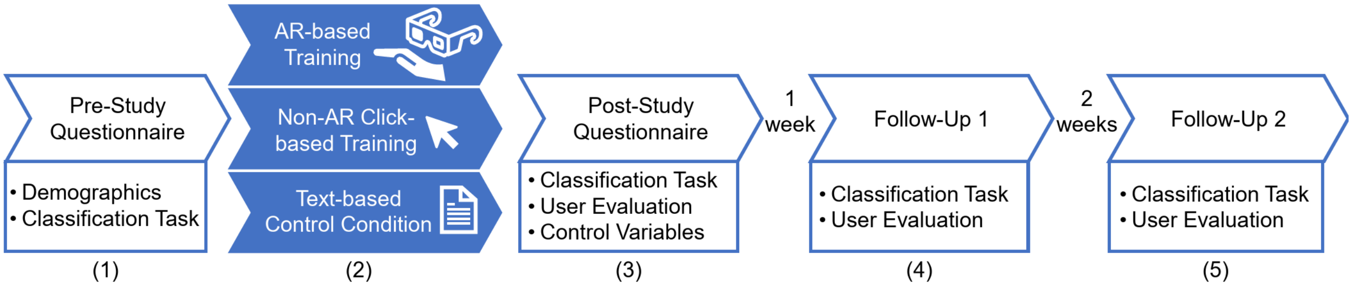
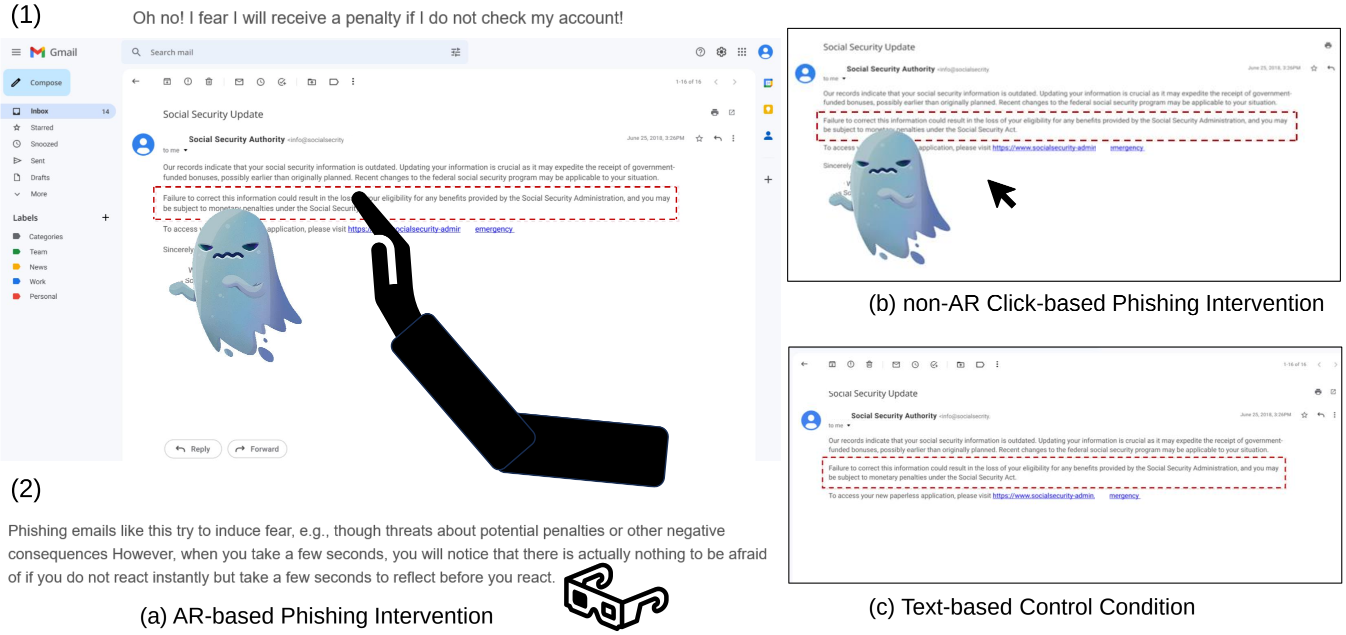


Figure 3: Visual summary of the study procedure.



Oh no! I fear I will receive a penalty if I do not check my account!

Social Security Update

Social Security Authority <info@socialsecrity

to me

June 25, 2018, 3:26PM

Our records indicate that your social security information is outdated. Updating your information is crucial as it may expedite the receipt of government-funded bonuses, possibly earlier than originally planned. Recent changes to the federal social security program may be applicable to your situation.

Failure to correct this information could result in the los our eligibility for any benefits provided by the Social Security Administration, and you may be subject to monetary penalties under the Social Securit

To access application, please visit https:/ ocialsecurity-admir emergency

Sincerely

Phishing emails like this try to induce fear, e.g., though threats about potential penalties or other negative consequences However, when you take a few seconds, you will notice that there is actually nothing to be afraid of if you do not react instantly but take a few seconds to reflect before you react.

(a) AR-based Phishing Intervention

(b) non-AR Click-based Phishing Intervention

Social Security Update

Social Security Authority <info@socialsecrity

to me

June 25, 2018, 3:26PM

Our records indicate that your social security information is outdated. Updating your information is crucial as it may expedite the receipt of government-funded bonuses, possibly earlier than originally planned. Recent changes to the federal social security program may be applicable to your situation.

Failure to correct this information could result in the loss of your eligibility for any benefits provided by the Social Security Administration, and you may be subject to monetary penalties under the Social Security Act.

To access your new paperless application, please visit https://www.socialsecurity-admin. mergency

(c) Text-based Control Condition

Figure 4: Visual comparison of the three intervention conditions.

for the AR-based condition to superimpose virtual content and enable gesture-based interaction. Participants in the click-based condition only interacted with the web app. The development and details are described in Section 3.2. Since the visibility of colors can vary between computer displays and AR environments (such as black being barely visible in the HoloLens 2), and because AR has a limited FOV, images and text in the AR condition were inverted or otherwise modified to make them more legible.

During the laboratory study, researchers were present for an initial introduction and final debriefing, as well as for resolving issues or assisting participants with the AR headsets. The researchers did not interact with participants, instead remotely observing participants from a separate room.

4.5 Eye Tracking

We tracked participants' eye movements throughout the laboratory study, except for participants in the AR-based condition while they were wearing the Hololens 2. We used this data for attention checks and analysis during the classification tasks. Eye tracking measures continuous data indicative of a user's visual attention [36], and serves as a good measure to enhance questionnaire data [57], making it possible to determine how much time and effort is spent on any specific task or visual area, and therefore how much cognitive processing they require.

We used Tobii 4C eye trackers [80] with a sample rate of 60 Hz. While no official metrics exist for this model, the comparable Tobii EyeX tracker possessed an accuracy of 0.6° and a precision of 0.25° [28]. The tracker was mounted at the bottom of the computer monitor, a Dell S2522HG with a resolution of 1920x1080 and a size of 24.5 inches. Participants were seated approximately 65 cm from the monitor. Using a 3-point calibration, we collected all eye tracking data intended for analysis within roughly 15 minutes.

4.6 Ethical Considerations

The study design followed established ethical guidelines for psychological research involving humans [5] and was approved by our university's ethics board. We minimised the potential for privacy invasion, e.g., by collecting age ranges instead of a concrete age and by factory resetting all devices used to collect personally identifiable data. The eye tracking data only contains coordinate points corresponding to the computer screen, but no images of faces or eyes. Prior to registering for the study, participants were already informed about the nature of the tasks. Before the study, participants were provided an informed consent sheet. Participation was voluntary and participants could abort the study and request the deletion of their data at any time without negative consequences. All participants received an equal payment. As there are regional minimum wage differences in our country, we compensated based on hourly wages for student assistants that exceed regional minimum wages to ensure fair compensation. Participants in the AR-based condition used a HoloLens 2 headset, which can lead to mild motion sickness after prolonged use [86], even though the HoloLens tries to minimise motion sickness, such as through low FOV [39]. As motion sickness generally only sets on after at least 20 to 30 minutes [39], we planned the AR training part to take no longer than 30 minutes. Furthermore, we excluded participants with a history of susceptibility to motion sickness, migraine, and fainting, as these factors contribute to motion sickness in AR [29].

4.7 Sample

A total of 120 participants took part in our study. They were recruited from a voluntary opt-in database associated with the university that contains people of different age groups and occupations. After the data collection, one participant was excluded due to technical problems and two more were excluded due to an apparent lack of seriousness in their responses, leaving us with a final sample of N=117 participants. A power analysis using G*Power 3.1 [19] estimated this amount of participants to be sufficient for detecting medium effects on phishing detection rates (Cohen's $d = 0.5$ and $\alpha = .05$) with a power of 0.98.

Of the participants, 56 identified as female, 61 as male, and none as diverse. The participants' age distribution was as follows: 59 were between 18-24, 40 between 25-34, 18 between 35-44. 86 participants have a university degree, 4 completed vocational school, and 27 participants completed secondary education. 13 participants indicated they had some background in cybersecurity, while 104 stated they did not. 110 participants stated they had never completed a cybersecurity training before, while five participants completed a training once, and two participants completed more than one training. The sample's affinity for technology interaction measured with the ATI scale ranging from "1 - completely disagree" to "6 - completely agree" was $M = 3.67$ ($SD = 0.86$), compared to an average score of 3.5 found in the general population [24].

5 User Study Results

In the following, we first detail our analysis method and then describe the findings regarding the training's impact on users. We describe changes to users' awareness for the cognitive biases, heuristics and norms exploited by attackers, as well as their phishing detection rates. Finally, we evaluate the users' evaluation of the intervention, and its design. For each measure, we first report the overall effects of the training answering RQ1 before comparing the interactive AR-based phishing intervention with the interactive click-based phishing intervention and the non-interactive text-based control condition to answer RQ2. As ATI is a useful grouping factor to evaluate proficiencies of coping with technology in both cybersecurity and AR research (e.g., [7, 24, 78]), we contrasted users whose ATI score fell below the median to those whose ATI score was above the median when analysing the interactive AR-based phishing intervention. This allowed us to broadly compare whether a particularly low or high technological affinity would affect the training effectiveness, similar to previous studies controlling for technological affinity in evaluating the effectiveness of AR tools (e.g., [7, 43]).

5.1 Analysis

For the quantitative analysis, we used repeated-measures ANOVAs to analyse significant differences between conditions over time, e.g., for comparing SUS scores. When the assumptions were violated, we instead employed a Kruskal-Wallis rank sum test. When measuring task performance, we adopted a mixed-effects regression to account for participant-specific factors and other variables that can affect phishing detection performance. We employed post-hoc tests to isolate specific effects between conditions, time points, or sub scales.

All qualitative analysis was conducted by two independent raters, who followed a deductive approach [49] to code mental model accuracy, interest, and engagement levels following pre-set categories with prototypical examples in a codebook. Initial inter-rater agreement across all these items was Cohen's $K = .88$, and any remaining disagreements were solved through discussion to assign the final code. For other open-ended responses such as suggestions or comments, two raters used open coding to inductively cluster the content into categories.

5.2 Users' Awareness for the Biases, Heuristics and Norms Exploited by Attackers

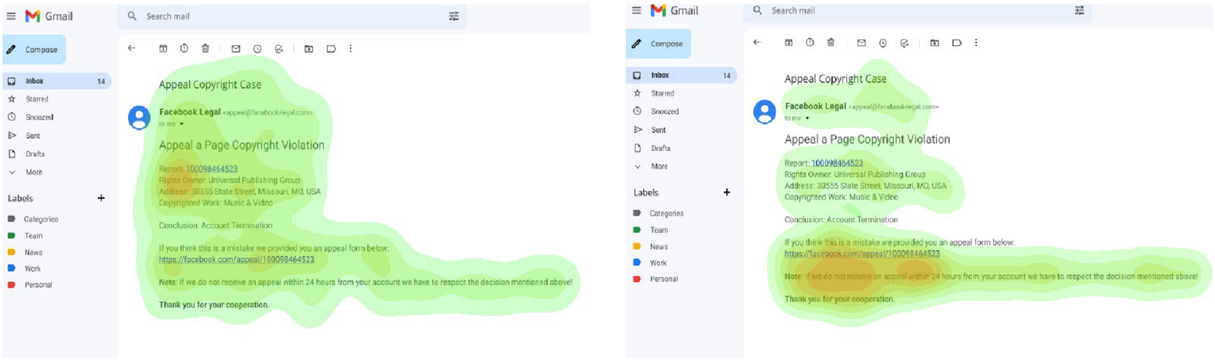
To account for the complexity of measuring awareness, we combined multiples measures for the purpose of triangulating, and to arrive at a more complete picture of users' changes in awareness levels. First, we used eye tracking as a behavioral measure and an indicator for visual attention. Second, we qualitatively captured the users' mental models of how phishers trick users to get detailed insights into how their awareness changes. Third, we made use of standardized quantitative scales such as the HAIS-Q. In the following, we detail the findings of all three measures.

5.2.1 *Visual Attention Captured through Eye Tracking.*

We measured all participants' eye movements during the pre- and post-intervention classification task in the laboratory to determine whether the training affected visual attention. Overall, the training led to a notable difference between pre- and post-intervention time points, demonstrated with an example in Fig. 5. However, there were no substantial differences between the training conditions.

5.2.2 *Self-Reported Awareness Captured through Mental Models of Phishing Attacks.*

We explored the participants' mental models of phishing attacks by asking them three times to describe their understanding of how phishing tricks people: once before the intervention, once after the intervention, and finally in the second



Appeal Copyright Case

Facebook Legal <appeal@facebook-legal.com>
to me

Appeal a Page Copyright Violation

Report: 100098464523
Rights Owner: Universal Publishing Group
Address: 30555 State Street, Missouri, MO, USA
Copyrighted Work: Music & Video

Conclusion: Account Termination

If you think this is a mistake we provided you an appeal form below:
https://facebook.com/appeal/100098464523

Note: If we do not receive an appeal within 24 hours from your account we have to respect the decision mentioned above!

Thank you for your cooperation.

Appeal Copyright Case

Facebook Legal <appeal@facebook-legal.com>
to me

Appeal a Page Copyright Violation

Report: 100098464523
Rights Owner: Universal Publishing Group
Address: 30555 State Street, Missouri, MO, USA
Copyrighted Work: Music & Video

Conclusion: Account Termination

If you think this is a mistake we provided you an appeal form below:
https://facebook.com/appeal/100098464523

Note: If we do not receive an appeal within 24 hours from your account we have to respect the decision mentioned above!

Thank you for your cooperation.

Pre Intervention

Post Intervention

Figure 5: Heatmap comparison of visual attention during the classification task between pre- and post-intervention time points.

follow-up survey. The textual answers were categorized in high, medium and low understanding following a deductive coding approach [49]. The results are shown in Table 3. An example for high understanding answers was *"They try to stress people with the fake deadlines, emergency situation, account loss, etc. Also, they can use a fake winning or end up sales or a complicated information which can be reached only in the attachment." (P24)*. Medium understanding answers contained phrases like *"to offer a fake prize, to redirect any unintentional click to a phishing website which is pretended to be an official website" (P99)*. Answers classified as low understanding contained statements such as *"get my IP related information, try to let me fill out their forms that ask for my information" (P81)*.

To answer RQ1 concerning the general effects of the training, the understanding improved substantially from pre- to post-intervention ($X^2(2) = 99.56, p < 0.001, \eta^2 = 0.28$), and remained high in the second follow-up survey. To better visualize these changes, we transformed the level of understanding into scores, where low understanding was equaled to -1, medium understanding equaled to 0, and good understanding equaled to 1, as can be seen in the rightmost column of Table 3. These averages cannot be interpreted numerically, but indicate whether understanding in the different training conditions tends more towards high or low understanding. However, these results contradict RQ2, as the AR-based condition shows comparable understanding to the control condition, and worse understanding when compared to the click-based condition.

When we excluded participants with an ATI below our sample's median of 3.5, understanding in the AR-based condition was much higher in both the post (0.88) and follow-up (0.59) time points, whereas they did not substantially change for the click-based condition (0.89 and 0.53, respectively).

5.2.3 Self-Reported Awareness Captured through Questionnaires. We measured users' cybersecurity awareness using the HAIS-Q email sub scale since our training specifically targeted phishing detection in emails. To not affect the interventions' effects, participants rated this scale only after the phishing intervention and in the two follow-up surveys. A comparison between post-intervention and follow-up 1 is visible in Table 8 in Appendix D. Overall, the scores dropped from the post-intervention to the follow-up studies ($X^2(2) = 32.48, p < 0.001, \eta^2 = 0.088$), yet becoming stable across the two follow-up surveys. We found a significant difference in cybersecurity awareness between the training conditions ($X^2(2) = 15.067, p < 0.001, \eta^2 = 0.038$), with a post-hoc pairwise comparison revealing a significantly higher score ($Z = 3.13, p = 0.003$) in the AR-based compared to the click-based condition. However, there were no significant differences between conditions in the follow-up studies.

Table 3: Overview on the number of participants categorized as having a low, medium, or high understanding. The average understanding of each group is expressed through a score.

Time	Group	Understanding (low/ medium/ high)	Score (Average)
Pre	Control	13 / 21 / 3	-0.27
	Interactive	15 / 24 / 3	-0.29
	AR	14 / 18 7 4	-0.28
	Total	42 / 63 / 10	-0.28
Post	Control	2 / 6 / 29	0.73
	Interactive	2 / 3 / 37	0.83
	AR	1 / 7 / 27	0.74
	Total	5 / 16 / 93	0.77
Follow-Up	Control	6 / 6 / 22	0.47
	Interactive	6 / 5 / 26	0.54
	AR	5 / 8 / 19	0.44
	Total	17 / 19 / 67	0.49

We also measured SEBIS scores, which showed no significant difference either between conditions ($F(2, 114) = 0.539, p = .585, \eta_p^2 = 0.009$) or time points ($F(2, 114) = 2.528, p = 0.115, \eta_p^2 = 0.022$).

5.3 Phishing Detection Ability

Similar to awareness, we measured the users' phishing detection ability in two different ways: 1) the objective detection rates captured through a classification task and 2) the subjective detection ability measured through self-report scales.

5.3.1 Objective Phishing Detection Rates. Participants were tasked to separate phishing from regular email in an email classification task before and after the intervention, as well as in both follow-up surveys. The performance in these tasks is summarized in Table 7 in Appendix D and visualised in Fig. 6. The intervention improved participants' phishing detection rate across all conditions, with the phishing detection accuracy increasing from .67 pre-intervention to .89 post-intervention and then remaining stable at .84 and .85 in both follow-up surveys. The two interactive conditions showed slightly higher improvements compared to the control condition. The highest post-intervention performance was seen in the AR-based condition with a phishing detection accuracy of .93. We calculated a mixed-effects regression model with phishing detection accuracy as the dependent variable and the training conditions, time points, and other control variables as predictors. This model revealed significant effects for all time points that were analysed using an ANOVA with $F(3, 275.35) = 73.91, p < .001, \eta_p^2 = 0.46$. However, there was no significant difference between post-intervention detection rates and follow-up 1 and 2 detection rates, indicating that training effects remain stable.

Further, there was a significant effect of the training condition on phishing detection accuracy. In the model, the click-based interactive ($\beta = 0.049, SE = 0.01, t(276) = 3.328, p = .001, d = 0.399$) and AR-based interactive conditions ($\beta = 0.030, SE = 0.01, t(278) = 2.077, p = .039, d = 0.244$) interacted significantly with the post-intervention detection rates, indicating that both affect improvements in phishing detection accuracy.

Overall, classification rates of non-phishing also improved after the training. Participants became more accurate at correctly identifying non-phishing emails as not being phishing emails. We again calculated a mixed-effects regression model with regular email detection accuracy as the dependent variable and conditions, time points, and other control variables as predictors. An ANOVA revealed a significant effect of overall time points ($F(3, 277.55) = 73.32, p < .001, \eta_p^2 = 0.46$). While the AR condition showed the highest pre-post phishing detection increase of 0.32 (compared to 0.20 in the click-based and 0.16 in the control condition), we found no significant difference between conditions ($F(2, 121.27) = 1.44, p = .241$) over all time points.

5.3.2 Subjective Phishing Detection Ability. Participants rated their confidence in phishing detection on three separate aspects: knowledge, ability, and alertness. These scores are summarised in Table 4. Scores for the click-based interactive condition were consistently the lowest, whereas the AR-based interactive condition saw the highest scores in ability and alertness confidence estimates. Overall, there was a significant effect of the training conditions for both ability ($X^2(2) = 11.81, p = .003, \eta^2 = 0.028$) and knowledge ($X^2(2) = 5.79, p = .048, \eta^2 = 0.011$), but not for alertness ($X^2(2) = 2.73, p = .256, \eta^2 = 0.002$).

In the two follow-up surveys, participants rated their confidence in detecting that phishing attacks try to trick them by abusing cognitive biases. These ratings are summarised in Table 9. While the AR-based condition showed the highest scores, there were no significant differences between training conditions.

5.4 User Evaluation

First, we were interested in the overall perceptions of the training, i.e., in terms of usability & user experience, cognitive effort and user engagement. We also collected open text responses about what users liked, disliked or what they would suggest. Second, we specifically explored the the users' evaluation of the condition-specific design elements such as the visual and auditory cues selected to represent certain biases, heuristics, and norms.

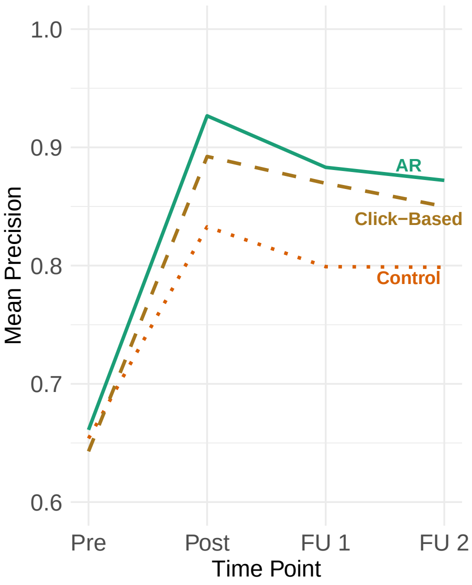


Figure 6: Visual overview of the precision to correctly classify phishing emails.

Table 4: Participants' Phishing Detection Confidence measured on a Scale from "1-very low" to "7-very high".

		Control M (SD)	Interactive M (SD)	AR M (SD)
Knowledge	Post	5.36 (0.79)	5.19 (1.33)	5.54 (1.05)
	FU1	5.39 (0.8)	5.29 (1.18)	5.67 (0.86)
	FU2	5.26 (0.9)	5 (1.14)	5.55 (1))
Ability	Post	5.41 (1.07)	4.91 (1.16)	5.53 (0.94)
	FU1	5.30 (1.07)	5.14 (1.27)	5.52 (0.75)
	FU2	5.57 (0.94)	5.07 (0.92)	5.60 (0.68)
Alertness	Post	5.46 (1.14)	5.41 (1.31)	5.78 (1.1)
	FU1	5.33 (1.18)	5.32 (1.22)	5.19 (1.21)
	FU2	5.60 (0.77)	5.19 (1)	5.65 (0.88)

5.4.1 Overall Evaluation of the Training.

Usability & User Experience. We measured usability and user engagement through the SUS [10] and the short form of the UES [58] scale, respectively. Both SUS and UES-SF scores are summarised in Table 10.

The control ($M = 78.08$) and interactive ($M = 74.05$) conditions received high SUS ratings, whereas the scores of the AR-based condition ($M = 50.35$) were significantly lower ($X^2(2) = 45.213, p < 0.001, \eta^2 = 0.379$). However, when we excluded participants with an ATI below our median of 3.5, the SUS scores for the remaining 18 participants in the AR-based condition were substantially higher ($M = 58.20$).

The overall UES-SF scores did not significantly differ between training conditions, whereas analysis of the sub dimensions revealed higher focused attention scores for both the click-based and AR-based conditions ($F(2, 114) = 3.74, p = .04, \eta^2 = 0.03$) but lower aesthetic appeal ($F(2, 114) = 1.87, p = .03, \eta^2 = 0.02$) and reward ($F(2, 114) = 2.11, p = .02, \eta^2 = 0.05$) scores for the AR-based condition.

Cognitive Effort. We also measured participants' task load during the training through the NASA-TLX scale [34]. An overview of the scores can be found in Table 11. We did not see significant differences between the three conditions on either overall scores nor on most sub dimensions. However, participants in the AR-based condition reported significantly lower perceived performance compared to the click-based condition ($X^2(2) = 6.8919, p = .032, \eta^2 = 0.043$).

Qualitative Analysis. We asked participants to list what they liked about the training and what suggestions they had in open-text fields. Two raters then categorized those answers and counted occurrences within categories. These findings are visualised in Fig. 7. The most liked aspect was the use of media, i.e., the illustrations and sounds, mentioned by 20 participants in the interactive click-based condition and 8 participants in the interactive AR-based condition. Interactivity was mentioned by 15 participants in the AR-based condition, while 15 participants in the control condition mentioned the use of phishing examples. Finally, 14 participants mentioned that they liked the clear and simple structure of the training, which included highlighting of important elements participants could focus on.

In the second open-text field, participants suggested the training should become even more interactive, either through gamification, an adventure-like training with multiple branches, or follow-up pages that make real-world implications clearer. Furthermore, they mentioned the need for more context information so that the information is easier to process, as well as additional focus on identifying unusual domains or links. Finally, 12 participants in the AR-based condition suggested the AR technology should be easier to use, and suggested a longer dedicated training, so that people are proficient with AR when they use it for the actual phishing intervention.

User Engagement. Finally, we asked participants to indicate whether the intervention affected their engagement or interest with cybersecurity in both follow-up surveys. Across all conditions, the majority of participants indicated that the intervention improved their engagement or interest in cybersecurity after 1 and 3 weeks. There were no significant differences between conditions. The corresponding textual answers were categorized into a clear, minor or no increase following a deductive coding approach [49]. The results are shown in Table 5.

Table 5: Overview of Increase in User Engagement with Cybersecurity or Phishing after the Training as Categorized from the Open Text Responses.

Time	Increase	Control	Interactive	AR	Total
FU1	No	8	5	2	15
	Minor	7	8	5	20
	Clear	12	15	14	41
FU2	No	7	5	4	16
	Minor	6	7	2	15
	Clear	17	14	14	45

5.4.2 *Evaluation of Condition-specific Design Elements.* We asked participants to rank each element and associated interactions by sorting them into categories: "like" for elements that were suitable representations, "dislike" for elements that were deemed unsuitable, and "neutral" for all other elements. A list of all elements can be seen in Table 2. The ranking of these elements is depicted in Fig. 8 in Appendix C. We did not collect this data from participants in the control condition since it did not contain these elements. Participants favoured urgency cues the most, with each element being placed in the like category by at least 60% of participants across both interactive conditions. Interactions representing the fear bias were more mixed, as the ghost element received high ratings but the spider element receiving low ratings, i.e., was deemed unsuitable as a representation for fear. This was particularly pronounced in the AR-based condition. Finally, complex elements and interactions that required more steps to complete received generally lower ratings in AR as compared to the click-based condition.

5.5 Summary of Findings

The phishing training led to improvements in phishing detection rates across all conditions. The AR-based intervention stood out with the highest post-intervention performance, achieving a phishing detection accuracy of .93. User evaluations indicated that the AR-based interactive condition significantly boosted confidence in phishing detection, particularly in terms of ability and alertness. Additionally, participants in the AR-based condition reported higher engagement levels and interest in cybersecurity. However, usability scores for the AR-based condition were substantially lower. We found this difference related to technological affinity, as people with higher ATI scores provided substantially higher usability scores. Overall, the AR-based intervention demonstrated promise in enhancing phishing detection and engagement, highlighting its potential in cybersecurity training.

6 Discussion

We first discuss the intervention effects, before exploring the intervention design and the potential of AR to enhance training. Finally, we describe limitations and highlight potential for future work.

6.1 Intervention Effects

Targeting cognitive biases seems to be effective in reducing phishing susceptibility as already the text-based training with no interaction led to substantial improvements in phishing detection. However, interactive training, either click-based or AR-based, was more effective compared to traditional text-based training. While the use of AR can enhance interventions, benefits from using it seem to depend on user-specific factors such as technological affinity and usability of the training. Therefore, the click-based intervention remains a viable alternative with strong training effects where an AR-based intervention is less suitable.

Phishing Detection. Our mixed-effects analysis reveals that users across all conditions exhibited consistent improvements in their

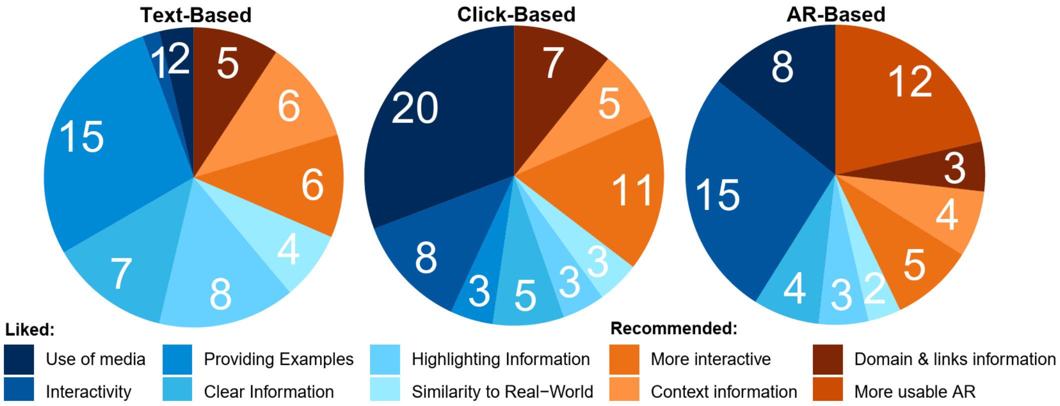


Figure 7: Summary of Participant's Most-Mentioned Feedback on the Intervention.

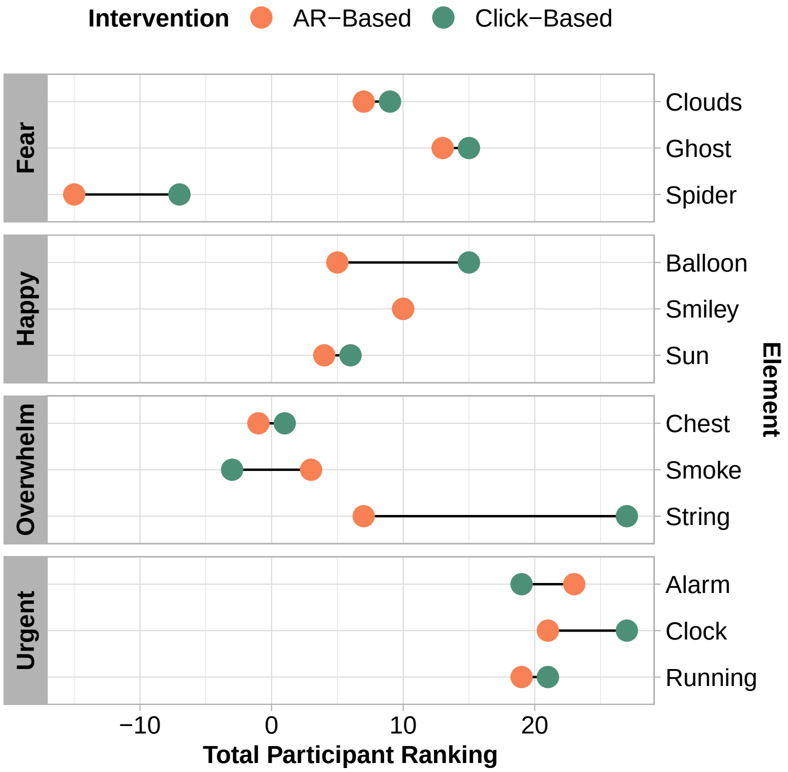


Figure 8: Visual comparison of the ranked intervention elements.

phishing detection behavior. This implies that there is a positive impact of the interventions on user behaviour, at least for three weeks after the intervention. This effect was substantially larger for the two interactive conditions as compared to the control condition, and even larger for the AR-based condition. Consistent with previous findings (e.g., [64]), this demonstrates the positive effect of increased interactivity on the effect of phishing interventions. Crucially, the participants' ability to correctly identify non-phishing email as not being phishing increased as well, thereby suggesting that the increases demonstrate actual improvements to participants' phishing detection ability, rather than just increasing suspicion towards emails in general. Furthermore, this increase was highest in the AR-based condition.

The phishing detection rates remained stable until the second follow-up survey 3 weeks after the laboratory study, although mean scores in the two follow-up surveys hinted at a slightly decreasing trend. This might be indicative of either a small general decline or a longer-term stable effect, with a short-lived initial phishing detection performance boost immediately after training.

Other Behavioural Changes. Our eye-tracking data offered insights into the impact of the interventions on visual attention. In particular, differences in visual attention between pre- and post-intervention classification tasks suggest that participants focus more on the email sections containing cognitive biases rather than reading the email in general. This result suggests that the performance improvements may be caused by stronger attention towards cognitive bias cues, which then lead to more accurate classification of the emails.

Understanding & Self-Estimates of Phishing Proficiency. Participants' understanding, especially of phishing tactics, notably increased following the training. These improvements were sustained in the second follow-up survey, indicating the long-term effectiveness of the training at enhancing user's phishing knowledge. Strikingly, when examining participants with high ATI scores above the median, we observed that the AR-based intervention had a more pronounced effect on increasing understanding compared to the click-based approach. These findings emphasize the potential of AR interventions in bolstering users' defenses against phishing attempts, but again reinforces that these effects are variable.

Participants' feedback also indicated that the interventions positively affected their engagement and interest in cybersecurity, with the majority reporting increased interest in the follow-up surveys. This implies that user behavior improvements extend beyond immediate detection to a broader impact on cybersecurity attitudes. The effect seemed particularly pronounced for the AR-based condition, where two thirds of participants describe a clear increase. This effect can likely be attributed to AR providing increased user engagement [44, 68, 91], which can translate to a higher interest in cybersecurity topics if the intervention itself is more engaging, thereby generalizing effects.

6.2 Intervention Design

While the primary goal was evaluating the feasibility and effectiveness of the training as a concept, we additionally measured how the design decisions affected users' perception of the training and its elements. Usability ratings indicated that both interactive and control conditions were well-received, while the AR-based condition's usability ratings were lower. Furthermore, participants in the AR-based condition reported lower perceived performance in the NASA-TLX questionnaire. This discrepancy was somewhat mitigated when participants with lower ATI scores were excluded. However, usability scores were still lower in the AR-based compared to the click-based condition, which is consistent with previous results showing that AR tools are perceived as less usable compared to desktop-based counterparts [62], likely influenced by a substantial difference in exposure between desktop-based and AR tools. Instead, the higher technological affinity of high ATI users might have helped users focus better on the learning experience, despite lower usability. This is consistent with previous findings, for instance, participants with higher ATI seem to have a higher preference for tangible compared to abstract interactions in AR [16] and are better able to use AR tools overall [7, 43].

In line with these findings, 12 participants suggested that the AR aspect of the AR-based training should be made easier to use, implying that they encountered some challenges. These findings indicate that the effect of the AR technology and the benefit people gain from using it is mediated by participant's proficiency in using it.

These challenges might provide an explanation for why we did not find significant differences between the AR-based and the interactive condition for many outcome variables. It might be that the AR-based condition led participants to perform better on average, but that this effect was weighed down by participants who were not able to effectively take advantage of the AR technology. This problem is not new, as the use of AR has been shown to lead to difficulties, especially for people with lower technical ability [26]. Future work should thus explore measures for AR proficiency or alternative extended reality technologies.

Representation & Interaction. The most effective intervention elements concerned urgency cues. On the other hand, cues invoking fear received mixed reception. While a spider element was consistently rated very low, a ghost element was rated much higher in terms of their suitability to represent fear. This indicates that people prefer a representation of the underlying cognitive bias and what it tries to induce, but do not want to be negatively affected by the representation itself.

Compared to ratings in the click-based condition, participants in the AR condition rated more complex elements and elements with a fearful association lower, whereas simple elements focusing on urgency were higher rated. The difference in participant's rating regarding complexity seems related to usability differences between the two conditions, where challenges relating to usability of AR were compounded by the need for more intricate interactions and more challenging perception. The lower rating of fearful elements, and especially low rating of the spider, is likely reinforced by the tendency of AR to tunnel attention towards objects, thereby making them more visceral [60, 62]. Perhaps also a stronger emotional response to the AR-based as compared to a 2D representation as found by Zhao et al. [103] might explain the difference in rating. Elements that are simple to understand and whose meaning can effectively be enhanced by the attention-catching nature of AR, such as the elements we used to represent urgency, seem to be most well-received by participants.

Striking the right balance between effective representation and avoiding negative emotional impacts is a crucial design consideration. Elements invoking negative or averse emotions might be less suitable for a cybersecurity training. On the other hand, simple elements whose representation manages to take advantage of the medium they're presented in would be most suitable, especially in an augmented reality setting. This interaction can be further affected by culture-specific or other differences. Therefore, careful consideration of training elements tailored to the specific use-case seems crucial.

Participants' feedback indicated that the use of media and degree of interactivity were well received and contributed to the intervention's acceptance. Importantly, we have to keep in mind that participants likely tend to mention elements that were most pronounced to them. For example, participants in the control condition might specifically highlight the use of phishing examples, whereas participants in the interactive conditions focus on describing the interactive, multi-media experience as that aspect was more visible to them and would be more quickly recalled.

6.3 Augmented Reality

The use of AR can enhance phishing interventions, by focusing attention, increasing user motivation to engage in the training, and higher long-term interest. Among the intervention groups, the AR-based condition demonstrated the highest post-intervention phishing detection accuracy of 0.93. This suggests that the immersive nature of AR may have a particularly pronounced impact on users' behavioural responses. Furthermore, participants in the AR-based condition scored well in other outcomes in comparison to the click-based condition, despite the intervention itself suffering from low usability scores compared to the other two conditions. For instance, they received higher HAIS-Q scores and reported more engagement and interest in cybersecurity following the intervention. This is consistent with previous findings showing higher engagement even with limited usability [62].

The benefits gained by using AR seem to be mediated by an affinity for using AR. This is most prominent in the low SUS ratings following the AR condition, which are substantially higher when controlled for technological affinity using the ATI. Participants' suggestions for improving the AR-based condition, including the need for a dedicated AR training session, emphasize the importance of ensuring user proficiency with AR technology. The positive impact and overall usefulness of AR interventions seems to be substantially influenced by user's familiarity and comfort with using AR.

The gestures and physical actions required for the AR-based condition also have the potential for providing inherent benefits themselves. For instance, it could be an effective way to integrate embodied learning into cybersecurity training [9, 38], to make it more interactive and an involved experience [30], or could even take on ceremonial aspects (cf. [65]). As argued by Goldman [30], cybersecurity training can become more interactive and exciting through "friction"—situations that require users to stop and refocus—particularly for training attempting to enhance systematic processing. Therefore, some of the effects presented in the results could potentially be replicated with other modalities as long as they involve similar interactions, e.g., through motion tracking. Different approaches might also be a promising way to bypass issues of low affinity as current AR devices are not yet mainstream.

To fully harness the potential of AR in phishing interventions, it may be essential to consider users' proficiency level and introduce appropriate familiarisation elements. These elements could be naturally integrated into the training itself as not to add additional costs, or more sophisticated training programs where higher familiarity is essential for an effective interaction with AR-based materials. Moreover, this shortcoming could be mitigated due to AR still becoming more prominent, with adoption rates projected to increase substantially in the following years [77]. As users become more accustomed to the technology and devices evolve with features like AI-based enhancements [68], the benefits of using AR should become even more pronounced. In a future where AR is pervasive, real-time interventions might be seamlessly integrated into environments with widespread AR adoption [33]. Nevertheless, the large-scale adoption of AR does not eliminate the need for training as technical detection of threats remains an open challenge.

6.4 Limitations and Future Work

Researchers face a dilemma between generalizability, precision in control and measurement, and realism of study context [50]. The complex intervention design and largely untested use of AR technology in cybersecurity required a controlled setting, thereby limiting generalizability. However, the setting allowed us to combine different sources of measurement, such as self-reports, detailed behavioural measures, and classification task data. Furthermore, our classification task did not exactly represent real-life conditions, but tried to emulate them. For this purpose, we intended to provide a wide range of realistic example emails as material. However, due to the difference in sources and the lack of a standardised and modern set of phishing emails, this led to noticeable variations in individual email's detection difficulty (see Table 6 in Appendix C). Additionally, introducing material that accurately reflects personal relevance pose further challenges, as creating tailored phishing scenarios remains a limitation in current phishing research. While we evaluated the persistence of effects in follow-up surveys, the data were collected within a month. These findings demonstrated a stable effect with some drop-offs, but more long-term observations over multiple months would provide additional insights as to how persistent effects are (cf. [64]). However, due to participant dropouts, a larger sample is necessary the longer the collection period is, which is particularly challenging with a complex laboratory study.

The training focused on cognitive biases and on fostering systematic thinking. This is useful on its own but can also complement training with a focus on technical details of phishing emails. We employed experts to establish common biases that could effectively be engaged with in training. However, while all experts where actively engaged in HCI and usable security research, two of them have not yet published a paper at the time of the workshop. Further research could investigate how cognitive bias training can enhance training that focuses on technical differences between phishing and legitimate email.

As there are only few AR-based training approaches in cybersecurity and even less for phishing available yet, comparability with other phishing interventions is limited. In addition, phishing study effects highly depend on the chosen email set for classification, further impairing their comparability (e.g., [84]). Therefore, we compared the AR-based approach against both a click-based condition to test the effect of AR, as well as a text-based control condition to test the effect of interactivity in combination with the cognitive bias training. We also included triangulating measures to evaluate the training, beyond an email classification task, i.e., security-related scales and control variables, behavioural measures, and subjective evaluations. Yet, the findings should be interpreted as indicators, rather than comparing specific outcomes like the phishing detection rates directly.

Our results revealed that the individual performance and perceived usability of AR-based training was influenced by the user's

AR proficiency. However, our sample size was not sufficiently large to allow for more detailed investigations. Still, the results suggest that user groups with low technological affinity may struggle when using an AR-based application. However, users with higher technological affinity may also possess other properties that increase intervention effectiveness, such as higher interest in cybersecurity topics. Further research could isolate these effects in more granular study designs. As our sample was somewhat biased towards young people (mostly between 18 and 35) with a slightly higher technical affinity than the average population, the effect might be even more pronounced for more representative samples including older people. Future research should explore ways to tailor AR interventions to different proficiency levels and to lower the initial threshold for novice users.

Finally, the training material used in this study represents an initial version used to test its feasibility. Future work should build on these findings to a) focus on the representations deemed suitable by participants and b) trial additional biases, representations, or interactions.

7 Conclusion

Sophisticated phishing attacks capitalise on human cognitive biases to bypass technical measures and deceive individuals, thus remaining a prevalent and evolving threat. Given the limitations of conventional user training methods, we investigated the potential of coupling AR technology with a human-centred training. We ran an expert design workshop, implemented a system for interactive AR training, and evaluated it in a user study with $N = 117$ participants.

The results show that interactive AR-based training significantly improves phishing detection rates and better addressed cognitive biases exploited by attackers than both the control and the non-AR interactive training. Furthermore, AR can enhance other outcomes such as higher cybersecurity awareness and interest. Still, a click-based training remains a viable alternative with benefits especially in cases where AR is not feasible. With the growing prevalence of AR and immersive technologies, this work is an important step towards designing effective human-centred security training of the future.

8 Data Availability Statement

The data and material that support the findings of this article are openly available in https://doi.org/10.3929/ethz-b-000721055.

Acknowledgments

This work is graciously supported by armasuisse Science and Technology. We thank Ryssa Moffat for her contributions to data visualisation.

References

[1] Sonam Adinolf, Peta Wyeth, Ross Brown, and Roger Altizer. 2020. Towards Designing Agent Based Virtual Reality Applications for Cybersecurity Training. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction (OzCHI '19)*. Association for Computing Machinery, New York, NY, USA, 452–456. doi:10.1145/3369457.3369515

[2] Nurul Akbar. 2014. *Analysing persuasion principles in phishing emails.* Master's thesis. University of Twente.

[3] Hamed Alqahtani and Manolya Kavakli-Thorne. 2020. Exploring Factors Affecting User's Cybersecurity Behaviour by Using Mobile Augmented Reality App (CybAR). In *Proceedings of the 2020 12th International Conference on Computer and Automation Engineering* (Sydney NSW Australia, 2020-02-14). ACM, New York, NY, USA, 129–135. doi:10.1145/3384613.3384629

[4] APWG. 2022. Phishing Activity Trends Report, 3rd Quarter 2022. https://docs.apwg.org/reports/apwg_trends_report_q3_2022.pdf

[5] American Psychological Association. 2023. Ethical Principles of Psychologists and Code of Conduct. https://www.apa.org/ethics/code, Accessed 30th August 2023.

[6] E. Z. Barsom, M. Graafland, and M. P. Schijven. 2016. Systematic review on the effectiveness of augmented reality applications in medical training. *Surgical Endoscopy* 30, 10 (2016), 4174–4183. doi:10.1007/s00464-016-4800-6

[7] Reinhard Bernsteiner, Christian Ploder, Thomas Dilger, Johannes Nigg, Teresa Spieß, and Rebecca Weichelt. 2021. An Analysis of an Augmented Reality Application to Support Service Staff in Industrial Maintenance. In *Knowledge Management in Organizations*, Lorna Uden, I-Hsien Ting, and Kai Wang (Eds.). Springer International Publishing, Cham, 457–467. doi:10.1007/978-3-030-81635-3_37

[8] Hugo Bijmans, Tim Booij, Anneke Schwedersky, Aria Nedgabat, and Rolf van Wegberg. 2021. Catching Phishers By Their Bait: Investigating the Dutch Phishing Landscape through Phishing Kit Detection. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, California, USA, 3757–3774. https://www.usenix.org/conference/usenixsecurity21/presentation/bijmans

[9] David Birchfield, Harvey Thornburg, M. Colleen Megowan-Romanowicz, Sarah Hatton, Brandon Mechtley, Igor Dolgov, and Winslow Burleson. 2008. Embodiment, Multimodality, and Composition: Convergent Themes across HCI and Education for Mixed-Reality Learning Environments. *Advances in Human-Computer Interaction* 2008, 1 (2008), 874563. doi:10.1155/2008/874563 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2008/874563.

[10] John Brooke. 1996. SUS: A 'Quick and Dirty' Usability Scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[11] Marcus Butavicius, Kathryn Parsons, Malcolm Pattinson, and Agata McCormac. 2016. Breaching the Human Firewall: Social engineering in Phishing and Spear-Phishing Emails. doi:10.48550/arXiv.1606.00887 arXiv:1606.00887 [cs]

[12] Casey Inez Canfield, Baruch Fischhoff, and Alex Davis. 2016. Quantifying Phishing Susceptibility for Detection and Behavior Decisions. *Human Factors* 58, 8 (Dec. 2016), 1158–1172. doi:10.1177/0018720816665025 Publisher: SAGE Publications Inc.

[13] Deanna D. Caputo, Shari Lawrence Pfleeger, Jesse D. Freeman, and M. Eric Johnson. 2014. Going Spear Phishing: Exploring Embedded Training and Awareness. *IEEE Security & Privacy* 12, 1 (2014), 28–38. doi:10.1109/MSP.2013.106 Conference Name: IEEE Security & Privacy.

[14] Hsin-Yi Chang, Theerapong Binali, Jyh-Chong Liang, Guo-Li Chiou, Kun-Hung Cheng, Silvia Wen-Yu Lee, and Chin-Chung Tsai. 2022. Ten years of augmented reality in education: A meta-analysis of (quasi-) experimental studies to investigate the impact. *Computers & Education* 191 (2022), 104641. doi:10.1016/j.compedu.2022.104641

[15] Yan-Ming Chiou, Chien-Chung Shen, Chrystalla Mouza, and Teomara Rutherford. 2021. Augmented Reality-Based Cybersecurity Education on Phishing. In *2021 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)* (2021-11). IEEE, New York, NY, USA, 228–231. doi:10.1109/AIVR52153.2021.00052

[16] Sarah Delgado Rodriguez, Priyasha Chatterjee, Anh Dao Phuong, Florian Alt, and Karola Marky. 2024. Do You Need to Touch? Exploring Correlations between Personal Attributes and Preferences for Tangible Privacy Mechanisms. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–23. doi:10.1145/3613904.3642863

[17] Amir Dirin and Teemu H. Laine. 2018. User Experience in Mobile Augmented Reality: Emotions, Challenges, Opportunities and Best Practices. *Computers* 7, 2 (2018), 33. doi:10.3390/computers7020033

[18] Serge Egelman and Eyal Peer. 2015. Scaling the Security Wall: Developing a Security Behavior Intentions Scale (SeBIS). In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (New York, NY, USA, 2015-04-18) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2873–2882. doi:10.1145/2702123.2702249

[19] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods* 41, 4 (2009), 1149–1160. doi:10.3758/BRM.41.4.1149

[20] Ana Ferreira, Lynne Coventry, and Gabriele Lenzini. 2015. Principles of persuasion in social engineering and their use in phishing. In *Human Aspects of Information Security, Privacy, and Trust: Third International Conference, HAS 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015. Proceedings 3*. Springer, Cham, Switzerland, 36–47.

[21] Ana Ferreira and Gabriele Lenzini. 2015. An analysis of social engineering principles in effective phishing. In *2015 Workshop on Socio-Technical Aspects in Security and Trust (STAST)* (2015-07-01). IEEE Computer Society, Washington,

D.C., United States, 9–16. doi:10.1109/STAST.2015.10

[22] OpenJS Foundation. 2023. Node.js JavaScript runtime environment. https://nodejs.org/

[23] Sender Policy Framework. 2006. Sender Policy Framework. http://www.openspf.org/

[24] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human–Computer Interaction* 35, 6 (2019), 456–467. doi:10.1080/10447318.2018.1456150 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2018.1456150.

[25] Edwin Donald Frauenstein and Stephen Flowerday. 2020. Susceptibility to phishing on social network sites: A personality information processing model. *Computers & Security* 94 (2020), 101862. doi:10.1016/j.cose.2020.101862

[26] Juan Garzón, Juan Pavón, and Silvia Baldiris. 2019. Systematic review and meta-analysis of augmented reality in educational settings. *Virtual Reality* 23, 4 (2019), 447–459. doi:10.1007/s10055-019-00379-9

[27] Nirit Gavish, Teresa Gutiérrez, Sabine Webel, Jorge Rodríguez, Matteo Peveri, Uli Bockholt, and Franco Tecchia. 2015. Evaluating virtual reality and augmented reality training for industrial maintenance and assembly tasks. *Interactive Learning Environments* 23, 6 (2015), 778–798. doi:10.1080/10494820.2013.815221 Publisher: Routledge _eprint: https://doi.org/10.1080/10494820.2013.815221.

[28] Agostino Gibaldi, Mauricio Vanegas, Peter J. Bex, and Guido Maiello. 2017. Evaluation of the Tobii EyeX Eye tracking controller and Matlab toolkit for research. *Behavior Research Methods* 49, 3 (June 2017), 923–946. doi:10.3758/s13428-016-0762-9

[29] John F. Golding, Aisha Rafiq, and Behrang Keshavarz. 2021. Predicting Individual Susceptibility to Visually Induced Motion Sickness by Questionnaire. *Frontiers in Virtual Reality* 2 (2021), 11 pages. https://www.frontiersin.org/articles/10.3389/frvir.2021.576871

[30] Eric H. Goldman. 2019. Push the Button: Making Security Training Fun and Interactive. *International Journal of Information Security and Cybercrime (IJISC)* 8, 1 (2019), 30–34. https://www.ceeol.com/search/article-detail?id=833997 Publisher: Asociatia Romana pentru Asigurarea Securitatii Informatiei.

[31] William J Gordon, Adam Wright, Robert J Glynn, Jigar Kadakia, Christina Mazzone, Elizabeth Leinbach, and Adam Landman. 2019. Evaluation of a mandatory phishing training program for high-risk employees at a US healthcare system. *Journal of the American Medical Informatics Association* 26, 6 (2019), 547–552. doi:10.1093/jamia/ocz005

[32] David Gragg. 2003. A multi-level defense against social engineering. *SANS Reading Room* 13 (2003), 1–21.

[33] Jens Grubert, Tobias Langlotz, Stefanie Zollmann, and Holger Regenbrecht. 2017. Towards Pervasive Augmented Reality: Context-Awareness in Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics* 23, 6 (June 2017), 1706–1724. doi:10.1109/TVCG.2016.2543720 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[34] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*, Peter A. Hancock and Najmedin Meshkati (Eds.). Human Mental Workload, Vol. 52. North-Holland, Amsterdam, Netherlands, 139–183. doi:10.1016/S0166-4115(08)62386-9

[35] Sumair Ijaz Hashmi, Niklas George, Eimaan Saqib, Fatima Ali, Nawaal Siddique, Shafay Kashif, Shahzaib Ali, Nida Ul Habib Bajwa, and Mobin Javed. 2023. Training Users to Recognize Persuasion Techniques in Vishing Calls. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2023-04-19) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, 1–8. doi:10.1145/3544549.3585823

[36] Kenneth Holmqvist, Marcus Nystrom, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van De Weijer. 2015. *Eye Tracking*. Oxford University Press, London, UK.

[37] Hang Hu, Peng Peng, and Gang Wang. 2018. Towards Understanding the Adoption of Anti-Spoofing Protocols in Email Systems. In *2018 IEEE Cybersecurity Development (SecDev)* (2018-09). IEEE, New York, NY, USA, 94–101. doi:10.1109/SecDev.2018.00020

[38] Gaoping Huang, Xun Qian, Tianyi Wang, Fagun Patel, Maitreya Sreeram, Yuanzhi Cao, Karthik Ramani, and Alexander J. Quinn. 2021. AdapTutAR: An Adaptive Tutoring System for Machine Tasks in Augmented Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3411764.3445283

[39] Claire L. Hughes, Cali Fidopiastis, Kay M. Stanney, Peyton S. Bailey, and Ernesto Ruiz. 2020. The Psychometrics of Cybersickness in Augmented Reality. *Frontiers in Virtual Reality* 1 (2020), 12 pages. https://www.frontiersin.org/articles/10.3389/frvir.2020.602954

[40] IBM. 2022. Cost of a Data Breach Report 2022. https://www.ibm.com/downloads/cas/3R8N1DZJ

[41] Daniel Jampen, Gürkan Gür, Thomas Sutter, and Bernhard Tellenbach. 2020. Don't click: towards an effective anti-phishing training. A comparative literature review. *Human-centric Computing and Information Sciences* 10, 1 (2020), 33. doi:10.1186/s13673-020-00237-7

[42] Jonah-Noël Kaiser, Thu Marianski, Frederike Jung, Mikołaj Woźniak, and Susanne Boll. 2022. Informed ShoppAR - Visualizing Privacy Information in Augmented Reality. In *Proceedings of Mensch und Computer 2022* (New York, NY, USA, 2022-09-15) *(MuC '22)*. Association for Computing Machinery, New York, NY, USA, 394–398. doi:10.1145/3543758.3549884

[43] Friedemann Kammler, Jonas Brinker, Jannis Vogel, Tahany Hmaid, and Oliver Thomas. 2019. How Do We Support Technical Tasks in the Age of Augmented Reality? Some Evidence from Prototyping in Mechanical Engineering. *ICIS 2019 Proceedings* 1 (Nov. 2019), 1–18. https://aisel.aisnet.org/icis2019/future_of_work/future_work/1

[44] B. Price Kerfoot and Nicole Kissane. 2014. The Use of Gamification to Boost Residents' Engagement in Simulation Training. *JAMA Surgery* 149, 11 (2014), 1208. doi:10.1001/jamasurg.2014.1779

[45] Ponnurangam Kumaraguru, Yong Rhee, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. 2007. Protecting People from Phishing: The Design and Evaluation of an Embedded Training Email System. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 905–914. doi:10.1145/1240624.1240760

[46] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. 2010. Teaching Johnny not to fall for phish. *ACM Transactions on Internet Technology* 10, 2 (2010), 1–31. doi:10.1145/1754393.1754396

[47] Eric Lin, Saul Greenberg, Eileah Trotter, David Ma, and John Aycock. 2011. Does domain highlighting help people identify phishing sites?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2011-05-07) *(CHI '11)*. ACM, New York, NY, USA, 2075–2084. doi:10.1145/1978942.1979244

[48] Xin Luo, Richard Brody, Alessandro Seazzu, and Stephen Burd. 2011. Social Engineering: The Neglected Human Factor for Information Security Management. *Information Resources Management Journal (IRMJ)* 24, 3 (2011), 1–8. doi:10.4018/irmj.2011070101 Publisher: IGI Global.

[49] Philipp Mayring. 2014. *Qualitative content analysis: theoretical foundation, basic procedures and software solution.* SSOAR, Mannheim, Germany.

[50] Joseph E McGrath. 1981. Dilemmatics: The study of research choices and dilemmas. *American Behavioral Scientist* 25, 2 (1981), 179–210.

[51] Microsoft. 2023. HoloLens2 - Overview, Features, and Specs. https://www.microsoft.com/en-us/hololens/hardware#document-experiences

[52] Microsoft. 2023. Mixed Reality Toolkit - Unity. https://github.com/microsoft/MixedRealityToolkit-Unity

[53] Mattia Mossano, Oksana Kulyk, Benjamin Maximillian Berens, Elena Marie Häußler, and Melanie Volkamer. 2023. Influence of URL Formatting on Users' Phishing URL Detection. In *Proceedings of the 2023 European Symposium on Usable Security*. Association for Computing Machinery, New York, NY, USA, 318. doi:10.1145/3617072.3617111

[54] Diep Nguyen and Gerrit Meixner. 2019. Gamified Augmented Reality Training for An Assembly Task: A Study About User Engagement. In *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)* (2019-09). IEEE, New York, NY, USA, 901–904. doi:10.15439/2019F136 ISSN: 2300-5963.

[55] Adam Oest, Yeganeh Safei, Adam Doupé, Gail-Joon Ahn, Brad Wardman, and Gary Warner. 2018. Inside a phisher's mind: Understanding the anti-phishing ecosystem through phishing kit analysis. In *2018 APWG Symposium on Electronic Crime Research (eCrime)* (2018-05). IEEE, New York, NY, USA, 1–12. doi:10.1109/ECRIME.2018.8376206 ISSN: 2159-1245.

[56] Daniela Oliveira, Harold Rocha, Huizi Yang, Donovan Ellis, Sandeep Dommaraju, Melis Muradoglu, Devon Weir, Adam Soliman, Tian Lin, and Natalie Ebner. 2017. Protecting people from phishing: the design and evaluation of an embedded training email system. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6412–6424. doi:10.1145/3025453.3025831

[57] Erica Olmsted-Hawala, Temika Holland, and Victor Quach. 2014. Usability Testing. In *Eye Tracking in User Experience Design*. Elsevier, Waltham, MA, USA, 49–80. doi:10.1016/B978-0-12-408138-3.00003-0

[58] Heather L. O'Brien, Paul Cairns, and Mark Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies* 112 (2018), 28–39. doi:10.1016/j.ijhcs.2018.01.004

[59] Christos Papakostas, Christos Troussas, Akrivi Krouska, and Cleo Sgouropoulou. 2021. Measuring User Experience, Usability and Interactivity of a Personalized Mobile Augmented Reality Training System. *Sensors* 21, 11 (2021), 3888. doi:10.3390/s21113888 Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

[60] Sangwon Park and Brigitte Stangl. 2020. Augmented reality experiences and sensation seeking. *Tourism Management* 77 (April 2020), 104023. doi:10.1016/j.tourman.2019.104023

[61] Kathryn Parsons, Dragana Calic, Malcolm Pattinson, Marcus Butavicius, Agata McCormac, and Tara Zwaans. 2017. The Human Aspects of Information Security

Questionnaire (HAIS-Q): Two further validation studies. *Computers & Security* 66 (2017), 40–51. doi:10.1016/j.cose.2017.01.004

[62] Iulian Radu. 2014. Augmented reality in education: a meta-review and cross-media analysis. *Personal and Ubiquitous Computing* 18, 6 (Aug. 2014), 1533–1543. doi:10.1007/s00779-013-0747-y

[63] Prashanth Rajivan and Cleotilde Gonzalez. 2018. Creative Persuasion: A Study on Adversarial Behaviors and Strategies in Phishing Attacks. *Frontiers in Psychology* 9 (2018), 14 pages. https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00135

[64] Benjamin Reinheimer, Lukas Aldag, Peter Mayer, Mattia Mossano, Reyhan Duezguen, Bettina Lofthouse, Tatiana von Landesberger, and Melanie Volkamer. 2020. An investigation of phishing awareness and education over time: When and how to best remind users. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*. Usenix, Berkeley, CA, USA, 259–284. https://www.usenix.org/conference/soups2020/presentation/reinheimer

[65] Karen Renaud and Marc Dupuis. 2023. Cybersecurity Insights Gleaned from World Religions. *Computers & Security* 132 (Sept. 2023), 103326. doi:10.1016/j.cose.2023.103326

[66] Joshua Reynolds, Deepak Kumar, Zane Ma, Rohan Subramanian, Meishan Wu, Martin Shelton, Joshua Mason, Emily Stark, and Michael Bailey. 2020. Measuring Identity Confusion with Uniform Resource Locators. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–12. doi:10.1145/3313831.3376298

[67] Giuseppe Riva, Rosa M. Baños, Cristina Botella, Fabrizia Mantovani, and Andrea Gaggioli. 2016. Transforming Experience: The Potential of Augmented Reality and Virtual Reality for Enhancing Personal and Clinical Change. *Frontiers in Psychiatry* 7 (2016), 14 pages. https://www.frontiersin.org/articles/10.3389/fpsyt.2016.00164

[68] Chandan K. Sahu, Crystal Young, and Rahul Rai. 2021. Artificial intelligence (AI) in augmented reality (AR)-assisted manufacturing applications: a review. *International Journal of Production Research* 59, 16 (2021), 4903–4959. doi:10.1080/00207543.2020.1859636

[69] M. A. Sasse, S. Brostoff, and D. Weirich. 2001. Transforming the 'Weakest Link' — a Human/Computer Interaction Approach to Usable and Effective Security. *BT Technology Journal* 19, 3 (2001), 122–131. doi:10.1023/A:1011902718709

[70] M. Angela Sasse, Jonas Hielscher, Jennifer Friedauer, and Annalina Buckmann. 2023. Rebooting IT security awareness–how organisations can encourage and sustain secure behaviours. In *Computer Security. ESORICS 2022 International Workshops*. Springer International Publishing, Cham, Switzerland, 248–265. doi:10.1007/978-3-031-25460-4_14

[71] Ryan Shandler and Miguel Alberto Gomez. 2022. The hidden threat of cyber-attacks – undermining public confidence in government. *Journal of Information Technology & Politics* 0, 0 (2022), 1–16. doi:10.1080/19331681.2022.2112796 Publisher: Routledge _eprint: https://doi.org/10.1080/19331681.2022.2112796.

[72] Filipo Sharevski, Jennifer Vander Loop, and Sarah Ferguson. 2024. "Oh, sh*t! I actually opened the document!": An Empirical Study of the Experiences with Suspicious Emails in Virtual Reality Headsets. doi:10.48550/arXiv.2412.01474 arXiv:2412.01474.

[73] Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. 2010. Who falls for phish? a demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2010-04-10). Association for Computing Machinery, New York, NY, USA, 373–382. doi:10.1145/1753326.1753383

[74] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason I. Hong, and Elizabeth Nunge. 2007. Anti-Phishing Phil: The Design and Evaluation of a Game that Teaches People Not to Fall for Phish. https://papers.ssrn.com/abstract=4248296

[75] Steve Sheng, Brad Wardman, Gary Warner, Lorrie Cranor, Jason Hong, and Chengshan Zhang. 2009. An Empirical Analysis of Phishing Blacklists. In *CEAS 2009 - Sixth Conference on Email and Anti-Spam* (2009-07-01). Carnegie Mellon University, Mountain View, California, USA, 10 pages. doi:10.1184/R1/6469805.v1

[76] Teodor Sommestad and Henrik Karlzén. 2019. A meta-analysis of field experiments on phishing susceptibility. In *2019 APWG Symposium on Electronic Crime Research (eCrime)* (2019-11). IEEE, New York, NY, USA, 1–14. doi:10.1109/eCrime47957.2019.9037502

[77] Statista. 2023. AR hardware B2C market users 2020-2027. https://www.statista.com/forecasts/1337381/ar-hardware-b2c-market-users-worldwide

[78] Stefan Sütterlin, Ricardo G. Lugo, Torvald F. Ask, Karl Veng, Jonathan Eck, Jonas Fritschi, Muhammed-Talha Özmen, Basil Bärreiter, and Benjamin J. Knox. 2022. The Role of IT Background for Metacognitive Accuracy, Confidence and Overestimation of Deep Fake Recognition Skills. In *Augmented Cognition*, Dylan D. Schmorrow and Cali M. Fidopiastis (Eds.). Springer International Publishing, Cham, 103–119. doi:10.1007/978-3-031-05457-0_9

[79] Jason Thomas. 2018. Individual Cyber Security: Empowering Employees to Resist Spear Phishing to Prevent Identity Theft and Ransomware Attacks. https://papers.ssrn.com/abstract=3171727

[80] Tobii. 2018. Tobii 4C.

[81] Unity. 2023. Unity Real-time Development Platform. https://unity.com/

[82] Tommy van Steen and Julia R.A. Deeleman. 2021. Successful Gamification of Cybersecurity Training. *Cyberpsychology, Behavior, and Social Networking* 24, 9 (2021), 593–598. doi:10.1089/cyber.2020.0526 Publisher: Mary Ann Liebert, Inc., publishers.

[83] Arun Vishwanath, Brynne Harrison, and Yu Jie Ng. 2018. Suspicion, Cognition, and Automaticity Model of Phishing Susceptibility. *Communication Research* 45, 8 (2018), 1146–1166. doi:10.1177/0093650215627483 Publisher: SAGE Publications Inc.

[84] Arun Vishwanath, Tejaswini Herath, Rui Chen, Jingguo Wang, and H. Raghav Rao. 2011. Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. *Decision Support Systems* 51, 3 (2011), 576–586. doi:10.1016/j.dss.2011.03.002

[85] Melanie Volkamer, Karen Renaud, and Paul Gerber. 2016. Spot the phish by checking the pruned URL. *Information & Computer Security* 24, 4 (Jan. 2016), 372–385. doi:10.1108/ICS-07-2015-0032 Publisher: Emerald Group Publishing Limited.

[86] Alla Vovk, Fridolin Wild, Will Guest, and Timo Kuula. 2018. Simulator Sickness in Augmented Reality Training Using the Microsoft HoloLens. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018-04-21) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–9. doi:10.1145/3173574.3173783

[87] Vue. 2023. Vue.js - The Progressive JavaScript Framework. https://vuejs.org/

[88] Zhuo Wang, Xiaoliang Bai, Shusheng Zhang, Mark Billinghurst, Weiping He, Peng Wang, Weiqi Lan, Haitao Min, and Yu Chen. 2022. A comprehensive review of augmented reality-based instruction in manual assembly, training and repair. *Robotics and Computer-Integrated Manufacturing* 78 (2022), 102407. doi:10.1016/j.rcim.2022.102407

[89] Rick Wash. 2020. How Experts Detect Phishing Scam Emails. *Proceedings of the ACM on Human-Computer Interaction* 4 (2020), 160:1–160:28. Issue CSCW2. doi:10.1145/3415231

[90] Rick Wash and Molly M. Cooper. 2018. Who Provides Phishing Training? Facts, Stories, and People Like Me. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2018-04-21) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3173574.3174066

[91] Yun Wen. 2021. Augmented reality enhanced cognitive engagement: designing classroom-based collaborative learning activities for young language learners. *Educational Technology Research and Development* 69, 2 (2021), 843–860. doi:10.1007/s11423-020-09893-z

[92] Zikai Alex Wen, Zhiqiu Lin, Rowena Chen, and Erik Andersen. 2019. What.Hack: Engaging Anti-Phishing Training Through a Role-playing Phishing Simulation Game. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019-05-02) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300338

[93] Brenda K. Wiederhold. 2021. Increasing Cybersecurity Through Emotional Engagement. *Cyberpsychology, Behavior and Social Networking* 24, 9 (2021), 579–580. doi:10.1089/cyber.2021.29224.editorial

[94] Christina Meilee Williams, Rahul Chaturvedi, and Krishnan Chakravarthy. 2020. Cybersecurity Risks in a Pandemic. *Journal of Medical Internet Research* 22, 9 (2020), e23692. doi:10.2196/23692

[95] Emma J. Williams, Amy Beardmore, and Adam N. Joinson. 2017. Individual differences in susceptibility to online influence: A theoretical review. *Computers in Human Behavior* 72 (July 2017), 412–421. doi:10.1016/j.chb.2017.03.002

[96] Emma J. Williams, Joanne Hinds, and Adam N. Joinson. 2018. Exploring susceptibility to phishing in the workplace. *International Journal of Human-Computer Studies* 120 (Dec. 2018), 1–13. doi:10.1016/j.ijhcs.2018.06.004

[97] Emma J. Williams and Danielle Polage. 2019. How persuasive is phishing email? The role of authentic design, influence and current events in email judgements. *Behaviour & Information Technology* 38, 2 (Feb. 2019), 184–197. doi:10.1080/0144929X.2018.1519599

[98] Eva Wolfangel. 2023. The Human Element in Cybercrime and Cybersecurity. CHI 2023 Opening Keynote. https://www.youtube.com/watch?v=LKUMRTLV49g

[99] Julia Woodward and Jaime Ruiz. 2023. Analytic Review of Using Augmented Reality for Situational Awareness. *IEEE transactions on visualization and computer graphics* 29, 4 (2023), 2166–2183. doi:10.1109/TVCG.2022.3141585

[100] Aiping Xiong, Robert W. Proctor, Weining Yang, and Ninghui Li. 2017. Is Domain Highlighting Actually Helpful in Identifying Phishing Web Pages? *Human Factors* 59, 4 (June 2017), 640–660. doi:10.1177/0018720816684064

[101] Shouhuai Xu. 2019. Cybersecurity Dynamics: A Foundation for the Science of Cybersecurity. In *Proactive and Dynamic Network Defense*, Cliff Wang and Zhuo Lu (Eds.). Springer International Publishing, Cham. Switzerland, 1–31. doi:10.1007/978-3-030-10597-6_1

[102] Weining Yang, Aiping Xiong, Jing Chen, Robert W. Proctor, and Ninghui Li. 2017. Use of Phishing Training to Improve Security Warning Compliance: Evidence from a Field Experiment. In *Proceedings of the Hot Topics in Science of Security:*

Symposium and Bootcamp (New York, NY, USA, 2017-04-04) *(HoTSoS)*. ACM, New York, NY, USA, 52–61. doi:10.1145/3055305.3055310

[103] Yuchen Zhao, Tulika Banerjee, Na Liu, and Jennifer G Kim. 2024. Grow With Me: Exploring the Integration of Augmented Reality and Health Tracking Technologies to Promote Physical Activity. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 182, 9 pages. doi:10.1145/3613905.3650907

[104] Olga Zielinska, Allaire Welk, Christopher B Mayhorn, and Emerson Murphy-Hill. 2016. The persuasive phish: Examining the social psychological principles hidden in phishing emails. In *Proceedings of the Symposium and Bootcamp on the Science of Security*. Association for Computing Machinery, New York, NY, USA, 126–126.

[105] Verena Zimmermann and Karen Renaud. 2019. Moving from a 'human-as-problem" to a 'human-as-solution" cybersecurity mindset. *International Journal of Human-Computer Studies* 131 (2019), 169–187. doi:10.1016/j.ijhcs.2019.05.005

A Questionnaires & Scales

A.1 Pre-Study Questionnaires

- Demographics
  - What is your age? [in age ranges]
  - What gender do you identify with? [male, female, non-binary/third gender, prefer not to say]
  - What is you highest level of education? [primary school, high school, professional education (e.g. commercial school), university degree (bachelor, master, PhD)]
  - Affinity for Technology Interaction Scale (ATI, [24])
- Cybersecurity & Phishing
  - Do you have a background in cybersecurity such as in education or occupation? [yes, no, other]
  - Security Behaviour Intention Scale (SeBIS, [18])
  - Have you ever participated in a phishing training? [Never, Once, More than once]
  - Please rate the following aspects: a) your knowledge about phishing emails, b) your ability to detect phishing emails, c) your alertedness to notice phishing emails in daily life [7-point Likert scale from "1-very low" to "7 - very high"]
  - How do you think phishing emails try to trick users (e.g., to click on a link, to provide credentials, or to download an attachment)? [open text field]
- Eye Tracking Calibration [Automatic calibration process]
- Email Classification task
  - Classification of 25 emails randomly drawn from a set of 50 emails
  - NASA Task Load Index (NASA-TLX, [34])

A.2 Post-Training Questionnaires

- Training Evaluation
  - How do you think phishing emails try to trick users (e.g., to click on a link, to provide credentials, or to download an attachment)? [open text field]
  - NASA Task Load Index (NASA-TLX, [34])
- User Evaluation
  - *Only for the interactive click-based and AR-based conditions:* In the following, you see symbols representing each training element. Please arrange them on the right side according to your preference. [Drag and Drop of 12 visual elements seen in the training in the three categories a) liked/suitable for illustrating the respective bias , b) neutral and c) disliked/ not suitable for illustrating the respective bias]
  - System Usability Scale (SUS, [10])
  - User Engagement Scale - Short Form (UES-SF, [58])
  - How interactive was the training in your opinion? [10-point Likert scale ranging from "1-not interactive at all" to "10- very interactive"]
  - How would you rate your overall experience with the training? [10-point Likert scale ranging from "1-extremely negative" to "10- extremely positive"]
  - What did you like about the training? [open text field]
  - Do you have any suggestions on how the training could be improved? [open text field]
  - Is there anything else you would like to mention? [open text field]
- Cybersecurity/ Phishing
  - How likely are you to detect that a phishing e-mail is trying to abuse a cognitive bias? [10-point Likert scale ranging from "1-very unlikely" to "10-very likely"]
  - Please rate the following aspects: a) your knowledge about phishing emails, b) your ability to detect phishing emails, c) your alertedness to notice phishing emails in daily life [7-point Likert scale from "1-very low" to "7 - very high"]
  - Security Behaviour Intention Scale (SeBIS, [18])
  - Human Aspects of Information Security Questionnaire (HAISQ, [61]) Subscale: Knowledge, Topic: E-Mail related behaviours [5-point Likert scale ranging from "1-strongly disagree" to "5 - strongly agree"]
- Eye Tracking Calibration [Automatic calibration process]
- Email Classification task
  - Classification of 25 emails randomly drawn from a set of 50 emails
  - NASA Task Load Index (NASA-TLX, [34])
- Are there any final comments you would like to submit? [open text field]

A.3 Follow-Up-Study Questionnaires

- How do you think phishing emails try to trick users (e.g., to click on a link, to provide credentials, or to download an attachment)? [open text field; only in second follow-up]
- How likely are you to detect that a phishing e-mail is trying to abuse a cognitive bias? [10-point Likert scale ranging from "1-very unlikely" to "10-very likely"]
- Please rate the following aspects: a) your knowledge about phishing emails, b) your ability to detect phishing emails, c) your alertedness to notice phishing emails in daily life [7-point Likert scale from "1-very low" to "7 - very high"]
- Human Aspects of Information Security Questionnaire (HAISQ, [61]), Subscale: Knowledge, Topic: E-Mail related behaviours [5-point Likert scale ranging from "1-strongly disagree" to "5 - strongly agree"]
- Did the training in the laboratory last week affect your interest or engagement with cybersecurity? [open text field]
- Classification of 24 emails, drawn in randomised order.
- NASA Task Load Index (NASA-TLX, [34]) [only in the second follow-up]

- Would you like to add any final comments? [open text field]

B Training Instructions

B.1 Introduction

What are cognitive biases?

The world is very complicated and we have to make a plethora of decisions every day. Therefore, we humans often rely on quick and automatic cognitive processes that help us make choices and to trigger certain actions. For example, humans often base decisions on emotions, social norms or on previous experiences. However, these automatic cognitive processes are based on general experiences and are not sensitive to context differences. Therefore, these automatic cognitive processes can sometimes be biased and lead to undesirable or incorrect decisions.

Phishing attacks try to exploit automatic cognitive processes and biases by creating a situation in which automatic reactions are triggered and lead us to act without thinking. For example, a phishing email might try to activate an emotion such as fear by pretending that our account has been hacked. This can result in unintended consequences such as clicking on a phishing link.

B.2 Instructions

Text-based control condition. In the following, you will undergo training where we specifically highlight how phishing attacks try to abuse automatic cognitive processes and biases with specific examples.

Please take your time to read through each example.

When you are ready, please proceed to start the training.

Interactive click-based condition. In the following, you will undergo training where we specifically highlight how phishing attacks try to abuse automatic cognitive processes and biases with specific examples.

In this training, you will see images of phishing emails over which we add annotations. Some of these annotations can be interacted with. We will also add additional information throughout the training.

Once all interactions with an image are complete, you can proceed to the next one. In the end, you will automatically be redirected back to this survey.

Please take your time to interact with each image.

Please put on your headset now, as there will be sound during the training.

When you are ready, please proceed to start the training.

Interactive AR-based condition. In the following, you will undergo training where we specifically highlight how phishing attacks try to abuse automatic cognitive processes and biases with specific examples:

- In this training, you will wear an augmented reality headset and work on the same computer as before.
- You will see images of phishing emails on the screen, over which we add annotations in augmented reality.
- You interact with these annotations using your hands.
- In addition, you will sometimes need to use the keyboard as will be indicated.

The goal is to first read the email. Afterward, you can interact with elements until you proceed to the next email.

Please continue to the next page now. When starting and when ending the training, a session supervisor will assist you.

B.3 Training Welcome Message

Welcome to the Training Module! During this session, you will be presented with several emails and asked to determine whether they are phishing attempts or not. To effectively identify a phishing email, pay attention to the small details within the message that may indicate it is not legitimate. We will provide explanations of the techniques used to recognize phishing emails after each example. Best of luck!

C Classification Task & Intervention Design

In Table 6, we list the emails in the classification task, where participants had to decide whether each was phishing or not. We additionally provide the pre-training phishing detection precision, which reveals that, on average, participants were already able to recognise the phishing emails as phishing prior to the training. Overall, 33% of benign emails and 30% of phishing emails did not contain a cognitive bias cue.

Table 6: Mean Precision Rate of Emails in the Classification Task Being Classified as Phishing. 0 Indicates No Classification as Phishing, While 1 Indicates Everyone Classified the Email as Phishing. We additionally list the type of cognitive bias the email was attempting to abuse.

Regular	Mean Rate	Cognitive Bias	Phishing	Mean Rate	Cognitive Bias
R01	.85	Familiarity	P01	.74	Fear
R02	.51	Happiness	P02	.47	None
R03	.58	Familiarity	P03	.92	Happiness
R04	.97	Happiness	P04	.53	None
R05	.11	None	P05	.60	Fear
R06	.04	Urgency	P06	.70	Authority
R07	.12	Happiness	P07	.61	Happiness
R08	.54	Happiness	P08	.35	Happiness
R09	.10	None	P09	.95	Urgency
R10	.69	Happiness	P10	.64	None
R11	.25	None	P11	.78	Fear
R12	.09	Fear	P12	.70	None
R13	.21	None	P13	.57	None
R14	.44	Urgency / Fear	P14	.88	Familiarity
R15	.14	Familiarity	P15	.70	Urgency
R16	.56	Fear	P16	.86	Happiness
R17	.27	Scarcity	P17	.75	Fear
R18	.24	None	P18	.68	Fear
R19	.19	None	P19	.32	Happiness
R20	.35	None	P20	.55	None
R21	.23	None			
R22	.34	None			
R23	.32	Familiarity			
R24	.29	Fear			
R25	.64	Fear			
R26	.19	Reciprocity			
R27	.25	Urgency			
R28	.27	Urgency			
R29	.08	None			
R30	.10	Familiarity			
Total	**.33**		**Total**	**.66**	

Table 8: Overview of HAIS-Q Email Sub Scale Scores, Each Comprising 3 Items and Ranging from 3 to 15 in Total.

Group	Time	Knowledge	Attitude	Behaviour
		M (SD)	M (SD)	M (SD)
Control	Post	11.15 (2.54)	13.59 (1.25)	12.15 (1.87)
	FU1	7.67 (5.45)	9.23 (6.33)	8.31 (5.92)
Interactive	Post	9.62 (2.80)	12.86 (1.82)	11.17 (2.26)
	FU1	6.81 (5.3)	8.64 (6.4)	7.45 (5.63)
AR	Post	11.11 (2.15)	13.19 (1.45)	12.14 (1.99)
	FU1	6.0 (5.38)	7.44 (6.45)	6.50 (5.76)
Total	Post	10.59 (2.61)	13.21 (1.55)	11.79 (2.09)
	FU1	6.85 (5.37)	8.47 (6.38)	7.44 (5.76)

D Detailed Results & Statistics

In this section, we provide additional details and summary tables of statistical tests.

Table 7: Phishing Detection Precision During Classification Task, Ranging from 0 Indicating No Correct Detection to 1, Indicating Perfect Detection

		Control	Interactive	AR
Precision	Pre	0.65	0.64	0.66
	Post	0.83	0.89	0.93
	Follow-Up 1	0.80	0.87	0.88
	Follow-Up 2	0.80	0.85	0.87

Table 9: Mean Confidence Ratings to Detect Cognitive Biases in Phishing Emails. Standard Deviations in Brackets.

Timepoint	Control	Interactive	AR	Total
Follow-Up 1	7.22 (1.25)	7.5 (1.29)	7.76 (1.55)	7.47 (1.35)
Follow-Up 2	7.5 (1.63)	7.19 (1.52)	7.4 (1.23)	7.36 (1.49)

Table 10: Comparison of Mean Usability Scores, Ranging from 0 as the Lowest Score to 100 as the Highest Score for SUS, and from 1 as the Lowest Score to 5 as the Highest Score for UES-SF. Standard Deviations in Brackets.

		Control	Interactive	AR	Total
SUS	Above median ATI	79.32 (9.17)	73.08 (12.12)	58.20 (16.11)	70.63 (16.36)
	All participants	78.08 (11.95)	74.05 (12.58)	50.35 (17.35)	67.81 (18.38)
UES-SF	Focused attention	2.79 (0.91)	3.04 (0.62)	3.04 (0.88)	3.96 (0.81)
	Aesthetic appeal	3.42 (0.78)	3.48 (0.89)	3.19 (1.13)	3.37 (0.94)
	Reward	3.84 (0.77)	3.87 (0.78)	3.54 (0.93)	3.76 (0.83)
	Total	3.35 (0.71)	3.45 (0.63)	3.25 (0.85)	3.36 (0.73)

Table 11: Mean NASA-TLX Ratings to Represent the Participant's Effort. Standard Deviations in Brackets.

Dimension	Condition	Pre	Training	Post
Mental	Control	54.1 (24.68)	38.59 (23.25)	55.26 (24.73)
	Interactive	57.86 (18.71)	42.14 (25.43)	53.69 (23.09)
	AR	41.94 (24.65)	36.11 (24.79)	42.36 (25.87)
Temporal	Control	42.82 (21.24)	32.56 (21.43)	42.44 (23.17)
	Interactive	43.33 (19.56)	35.24 (21.86)	42.86 (21.13)
	AR	31.94 (18.95)	25.42 (18.45)	33.61 (20.79)
Performance	Control	54.62 (20.34)	73.08 (18.2)	67.05 (16.61)
	Interactive	49.52 (16.63)	77.5 (14.24)	64.52 (18.67)
	AR	58.61 (22.06)	61.67 (27.44)	67.22 (15.6)
Effort	Control	58.21 (21.93)	42.44 (21.36)	51.28 (23.86)
	Interactive	59.29 (17.09)	42.86 (24.25)	51.07 (21.93)
	AR	48.75 (22.21)	43.89 (25.72)	45.97 (24.14)
Frustration	Control	53.72 (26.1)	31.15 (24.48)	42.44 (23.67)
	Interactive	51.67 (24.78)	29.05 (25.67)	40.12 (26.63)
	AR	36.67 (25.86)	41.39 (27.64)	36.11 (26.38)