Phishing to improve detection

Sarah Y. Zheng
University College London (UCL)
London, UK
sarah.zheng.16@ucl.ac.uk

Ingolf Becker
University College London (UCL)
London, UK
i.becker@ucl.ac.uk

ABSTRACT

Phishing e-mail scams continue to threaten organisations around the world. With generative artificial intelligence, conventional phishing detection advice such as looking out for linguistic errors and bad layouts will become obsolete. New approaches to improve people's ability to detect phishing are essential. We report on promising results from two experiments (total $N = 183$) that engaging people with an adversarial mindset improves their ability to detect phishing e-mails compared to those who received conventional or no training. Participants who completed conventional training were nearly three times as likely to fall for a simulated phishing attack compared to those who completed the adversarial training, in which they watched a fictitious cybercriminal explain how to devise a targeted phishing e-mail, and then wrote targeted phishing e-mails themselves. Although further research is needed to examine the training's long-term efficacy with larger sample sizes, the present findings show an encouraging alternative to conventional phishing training approaches.

CCS CONCEPTS

• **Security and privacy → Usability in security and privacy**.

KEYWORDS

phishing detection, cybersecurity training, adversarial mindset

ACM Reference Format:
Sarah Y. Zheng and Ingolf Becker. 2023. Phishing to improve detection. In *The 2023 European Symposium on Usable Security (EuroUSEC 2023), October 16–17, 2023, Copenhagen, Denmark.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3617072.3617121

1 INTRODUCTION

The threat of online scams has steadily increased since the COVID-19 pandemic [29, 34, 22, 8] and is set to become more sophisticated with the use of generative artificial intelligence (AI). For instance, industry reports observed a growth of linguistically more complex phishing e-mail scams since the release of ChatGPT [48]. This suggests that generative AI, capable of producing content indistinguishable from human creations [35, 18], is already being used to generate phishing e-mail scams.[1] As a result, conventional phishing detection advice such as looking out for grammar and spelling mistakes and bad layouts [33] will likely become less effective. Together with findings that suggest a limited efficacy of conventional phishing awareness training [42, 33, 24, 28, 55], it is therefore imperative to develop better ways to help people detect phishing e-mail scams. Here, we propose an adversarial training concept to improve people's ability to detect phishing e-mail scams, and compare its efficacy to those without additional training (Experiment 1) and those who completed a conventional phishing awareness training (Experiment 2).

The notion of "thinking like an attacker" (i.e., an adversarial mindset) is not new to cybersecurity practitioners, but has rarely been researched in the context of educating the general public. Only one previous study found that letting people create phishing URLs themselves led to better detection of phishing URLs [37]. As most people are mostly honest themselves [45, 13] and scammers are said to exploit people's trusting nature, engaging with an adversarial mindset may particularly help people with understanding what tactics scammers use and why. In other words, once people can imagine what a scammer may say, it may be easier for them to detect them. In line with this phenomenon of "self-projection" to understand others [1, 4, 6, 9, 23, 30, 31, 40], we developed an adversarial training video in which a cybercriminal explains how they would craft a phishing e-mail highly targeted at a single individual (i.e., a spearphishing e-mail) and then ask people to write three phishing e-mails themselves. Indeed, we find promising results that such adversarial training reduces phishing susceptibility compared to people who received no additional (Experiment 1) or conventional training (Experiment 2).

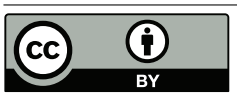
Next, we will turn to prior approaches to phishing detection education (Section 2), followed by the methods (Section 3) and results (Section 4) of our experiments. We conclude with a reflection on the ethics, limitations and future feasibility of our adversarial training concept, with the speculation that the concept may well apply across scam contexts (Section 5).

2 BACKGROUND

Methods to improve people's ability to detect phishing heavily rely on training and education programs [14, 19]. They are often part of larger organisational cybersecurity training efforts, including simulated phishing tests that present additional phishing training after people click on simulated phishing links [49]. While the contents of these training programs are usually comparable, various delivery approaches have been tried. Prior works range from conventional training programs in text- and/or video-based online or in-person formats [7, 20, 44, 51, 53] and serious games [26, 27, 15, 11, 5, 46, 52, 17], to simulated phishing campaigns [26, 3, 49, 25, 10].

To see the long-term efficacy of a conventional type of phishing training, Reinheimer et al. analysed the phishing detection rates of 409 public sector employees from four months after they completed

[1] Note that we regard phishing as a quintessential type of online scam and distinguish phishing from spam e-mails by the former's malicious nature.



EuroUSEC 2023, October 16–17, 2023, Copenhagen, Denmark

ACM ISBN 979-8-4007-0814-5/23/10.
https://doi.org/10.1145/3617072.3617121

the training. They found that the training effects on people's knowledge waned from six months after completing it, suggesting that people need to be reminded of training content every six months for training to remain effective [42]. In terms of what training reminder format works best, they found that video and interactive phishing e-mail examples work best [42]. Thus, while conventional phishing education may increase people's awareness of phishing and detection abilities in the short term, many of them may need more frequent reminders than often is the case to remain effective [42].

Serious games "Anti-Phishing Phil" [27] and "PHISHY" [15] have initially shown promising results of engaging people in a new and fun way to learn how to recognise phishing e-mails, but their efficacy compared to conventional training, nor their long-term efficacy has been researched or published yet. The use of simulated phishing attacks has in turn been criticised for its burdensome treatment of employees [49]. Moreover, results from a recent large-scale study suggest that simulated attacks with training after people fall for them has a worsening effect on their phishing susceptibility [28]. This is possibly due to a false sense of security induced by the simulations and embedded training.

The limited efficacy of these training approaches may also be due to the lack of usability of often given cybersecurity advice. Redmiles et al. surveyed the practical use of 1264 online privacy and security recommendations and found that the majority of advice was (indeed) not fully understandable for a general public [41]. In line with this, most conventional phishing training advises people to hover over links found in e-mails before clicking them to see if they may be malicious [33], which may be especially unhelpful since people often wrongly interpret URLs [2].

A few studies tested alternative approaches to phishing detection training that, for example, place less emphasis on technical details. Jensen et al. hypothesised that people may be more susceptible to phishing e-mails when they are rushed and busy, and thus proposed a mindfulness-based phishing awareness training that taught individuals to stop and reflect on the plausibility of requests made in e-mails. This approach diverged from giving conventional advice and effectively reduced phishing susceptibility, though mainly for those who had prior knowledge about phishing [21].

Wash and Cooper tested the effect of "teachers" and content format: how effective stories told by peers would be compared to conventional phishing detection advice provided by security experts. They found that stories-based phishing training was more effective when told by peers and "facts-and-advice"-based phishing training led to better phishing detection when told by experts, and that the conventional advice from experts was most effective [51]. These results suggest that people have strong expectations regarding how peers and security experts talk about security matters, which in turn affects training efficacy.

Collectively, these works imply the need to develop new approaches to make phishing education more effective. Here, we test an approach rooted in the psychology of judgement and decision-making, where we rely on the premises that most people (i) are honest [45, 13] and do not engage in phishing scams themselves and (ii) use their own behaviour to infer the honesty of others [1, 4, 6, 9, 23, 30, 31, 40]. As a result, honest people may be particularly prone to phishing scams, as the thought of someone else trying to scam them may not occur to them in the first place. Thus, by engaging people with an adversarial mindset, we aim to improve their understanding of how and why scammers may use certain phishing tactics, which we expect to enhance people's detection ability.

This approach inherently departs from teaching users how to spot phishing e-mails based on, e.g. inspecting URLs, but indirectly encourages people to think of e-mail contents that are most prone to be used in phishing scams. We take the role of (perceived) training sources into account in the creation of our adversarial and conventional phishing training, where the former involves an actor dressed as a louche cybercriminal and the latter the same actor dressed as a professional security adviser.

3 METHODS

We performed two experiments to measure the efficacy of an adversarial approach to phishing education. Experiment 1 was conducted with online participants as a pilot study to compare the adversarial training with the phishing detection of people without additional training. Experiment 2 was conducted to provide a more ecologically valid test with e-mail users from a large public sector partner organisation, that compared the adversarial training's efficacy to that of a conventional phishing awareness training. Both studies received prior ethics approval from our departmental research ethics board. See Section 5.2 for a more comprehensive discussion of the ethics of this study.

3.1 Participants Experiment 1

We recruited 20 participants for the adversarial training group and 24 participants for the control group (i.e., no training) through the Prolific crowd-sourcing platform. The adversarial training group first completed the training task and were selected to participate in an e-mail processing task two weeks later. We did not make them aware of the two-stage nature of the study to avoid biasing their responses toward phishing detection in the second task. Fifteen out of 20 adversarial training group participants also completed the second task and could therefore be analysed. The control group only completed the "second" e-mail processing task. Experiment 1 was kept small intentionally due to budget constraints and was therefore set up as a pilot to inform the pursuit of Experiment 2. All participants were compensated at a rate of £9 per hour and each task took approximately 30 minutes to complete.

3.2 Participants Experiment 2

Over the course of seven months, we recruited 104 participants for the conventional training and 40 participants for the adversarial training through e-mail invitations sent to all our partner organisation's staff, and distributed the study invitation across 35 internal departments and their internal research volunteers pool. Due to organisational constraints, we could not include a third "no training" group. To encourage participation, all participants were entered into a draw to win an iPad. They were informed that the study would not take more than 30 minutes to complete and aimed to gather feedback to help improve phishing education materials. Participants were not made aware of the two different testing conditions. After clicking the provided link to the training task, they were randomly assigned to either of two training conditions. We

only recruited from one organisation to be able to use simulated phishing attacks to measure phishing detection in an ecologically valid way.

3.3 Phishing detection training task

To test the potential of engaging people with an adversarial mindset to improve phishing detection, we developed an adversarial and conventional phishing awareness training. After participants gave informed consent and navigated to the next page, the training video started playing automatically. They were allowed to restart, play and pause the video at any point. Skipping to the end of the video was disabled. When the video finished playing, a button appeared below it to proceed to the next part of the training task.

3.3.1 Adversarial mindset training. The adversarial training video showed a professional male actor dressed in a black hoodie and sunglasses with the hood on at a desk in a dark computer servers room, to look like a stereotypical cybercriminal (see Figure 1). The actor used various jokes throughout the video, meant to actively engage the viewer with the adversary's thought process. In the video, he talked the viewer through each step in devising a realistic spearphishing e-mail, based on information commonly available through social media (here, LinkedIn). It started with finding a public cybersecurity conference on LinkedIn. He then browsed the event's attendants and found the CEO of one of the event's organising companies among the attendants list as well as their executive assistant. To impersonate the CEO's e-mail, the cybercriminal created a GMail address that followed the company's e-mail naming convention. The last part of the video shows him typing a spearphishing e-mail to the executive assistant, requesting the assistant to urgently buy Amazon vouchers as a gift for their event's guest speakers. See Figure 1.

After watching the video, participants were asked to describe in one word how they felt about the video, and rate how engaging and useful they found the video on a slider scale from "not engaging/useful at all" to "very engaging/useful". Then, they were informed that the next part of the study would require them to write three spearphishing e-mails themselves according to three scenarios. If they felt uncomfortable with proceeding to this part of the task, they were instructed to discontinue their participation (i.e., return to Prolific in Experiment 1 or close the task in Experiment 2) and excluded from any analyses. Otherwise, they clicked "Next".

Three phishing scenarios were presented one at a time, each containing a brief explanation of the e-mail context, the goal of the to be written e-mail, a hint to think of what fake e-mail address to use, and a target recipient's name, organisation and e-mail address. Participants were then asked to come up with a sender name, sender e-mail address and write a convincing e-mail message. The scenarios were: (i) a Nigerian prince scam aimed at convincing the recipient to transfer money as soon as they can, (ii) a credential harvest attack aimed at a senior employee at the United Nations, (iii) an urgent request to an event organiser to buy Amazon gift vouchers to reward guest speakers. After completing all e-mails, participants were asked to describe any strategy they used in writing the e-mails, if they had any concerns about them and rated how difficult they found the task on a slider scale from "very easy" to "very difficult".

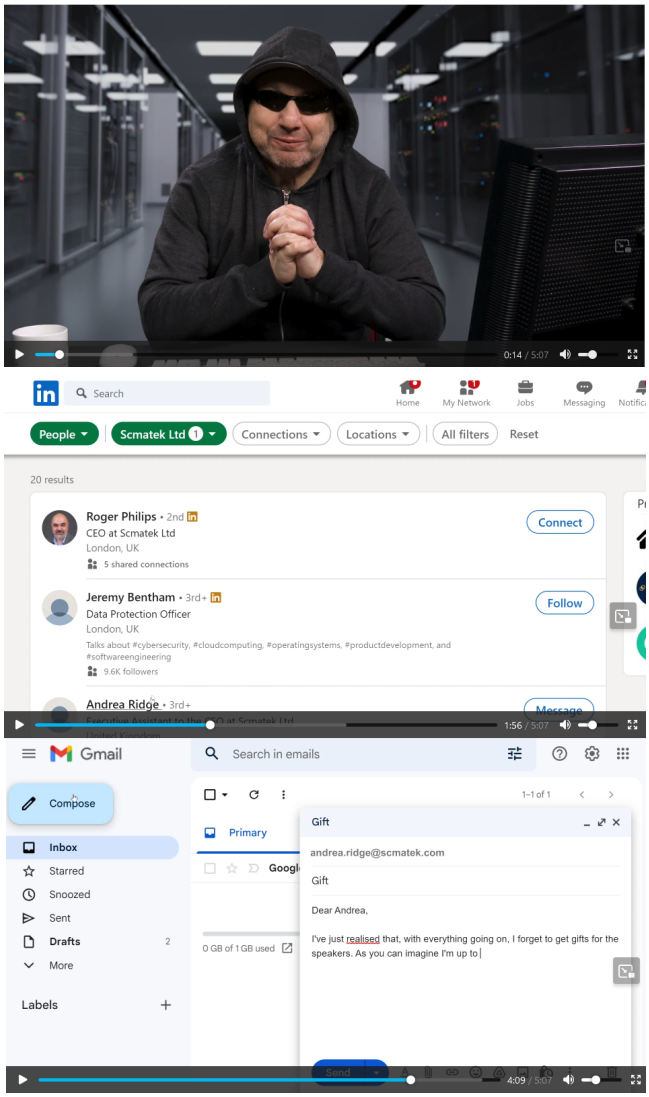


Figure 1: Stills from adversarial mindset training video.

After this, participants answered demographics questions (age, gender, education level, occupation status), how often they receive phishing e-mails themselves (daily, weekly, monthly, rarely), how many cybersecurity or anti-phishing training they completed before, rated how knowledgeable they are about cybersecurity on a 6-point scale from "not at all knowledgeable" to "very knowledgeable", and provided optional feedback on the study. Finally, participants were debriefed to not pursue targeted phishing attacks in real life, that none of their e-mails would be used in real attacks and that the study's sole aim is to educate them about how to detect phishing e-mail scams.

3.3.2 Conventional training. The conventional training consisted of a video on phishing detection, in which the same professional male actor, dressed in a smart casual suit, sat at a desk against the backdrop of a well-lit office space (see Figure 2). He explained

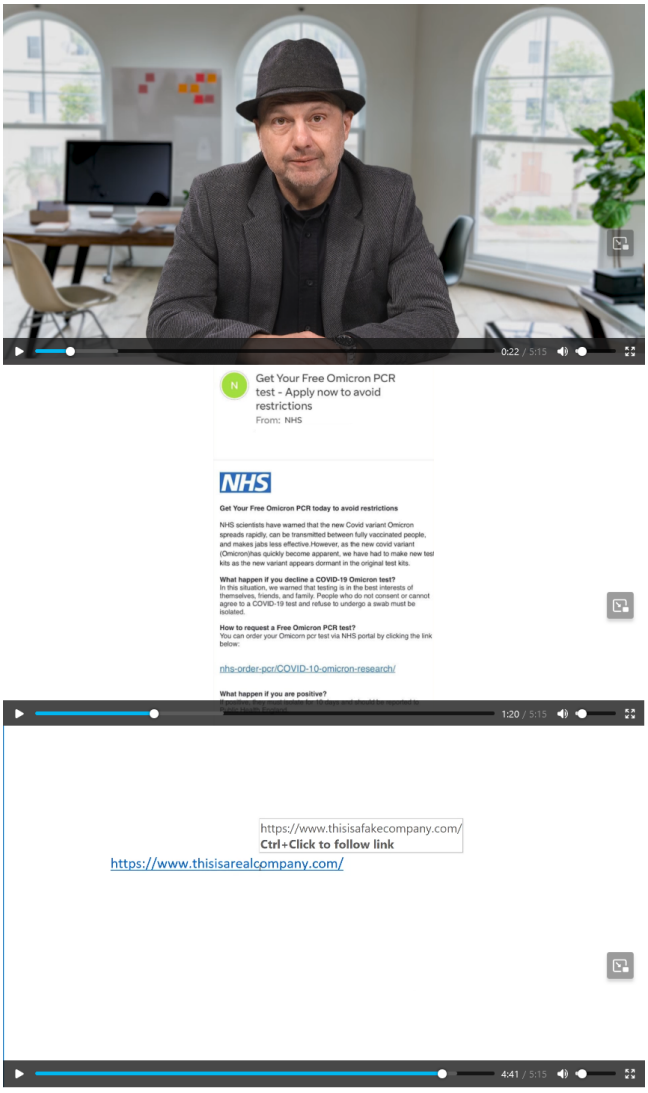


Figure 2: Stills from conventional phishing awareness training video.

what phishing e-mails are and how to detect them. He showed screenshots of two real phishing examples purporting to be from the NHS—taken from real attacks observed during the COVID-19 pandemic, and a LinkedIn profile and spearphishing e-mail to the person in the profile that urgently requested them to purchase gift vouchers for presenters at an event they organised. The latter was to have a comparable example to the one used in the adversarial training video. Whilst describing the examples, the actor pointed out the various suspicious signs to look out for that may indicate phishing, e.g. how to verify URLs and sender details, in line with often given advice in cybersecurity education [41]. The actor's tone of voice remained serious and professional throughout the video to give viewers the impression of an actual security adviser. See Figure 2.

After the video, participants were asked to describe in one word how they felt about the video, and rate how engaging and useful they found the video on a sliding scale from "not engaging/useful at all" to "very engaging/useful". Next, they answered the same demographics and experiential questions as the adversarial training group. In the version for Experiment 2, we also asked for participants' anonymous e-mail alias and department.

3.4 Phishing detection measurement

We measured people's phishing e-mail detection ability through an online e-mail processing task (Experiment 1) or simulated phishing tests (Experiment 2) two weeks after they completed the training task. Due to the nature of these detection methods, the accuracy metrics differ per experiment. In Experiment 1, we used independent samples t-tests to compare overall performances between the adversarial and no training groups. In Experiment 2, we used Fisher's exact test to test for equality of proportions of participants who fell for the simulated attacks in the adversarial versus conventional training groups, as well as logistic regression analyses to see the effect of testing condition on phishing victimisation while controlling for other factors (e.g. demographics and prior training). Statistical significance was evaluated against a 0.05 threshold. We also report the effect sizes for all statistical tests related to phishing detection and performed post hoc power ($1 - \beta$) analyses for the main effects of interest (phishing detection recall and precision in Experiment 1 and victimisation proportions in Experiment 2). All analyses were performed in R [12].

3.4.1 Experiment 1: Mock inbox task. Since we were not allowed to ask for Prolific participants' e-mail addresses to send them simulated phishing e-mails, we developed a basic Outlook e-mail interface to simulate an e-mail processing experience. Participants were told the study aimed to understand how people process e-mails. They were asked to play the role of an executive director at a fictitious corporation, and process their e-mails in the simulated Outlook inbox. Each e-mail could be replied to, forwarded, deleted or kept in the inbox. After selecting an action, participants specified their reason for doing so in a secondary bar menu that appeared beneath the main actions bar. For example, if participants clicked "delete", they had to select one of the following reasons: "uninteresting or irrelevant", "no action required", "spam", "phishing" or "other". All reasons were based on qualitative works on how people process e-mails [54, 50, 38]. Phishing e-mails were said to be detected if participants "deleted" the e-mail and selected "Phishing" as the reason. Phishing e-mails that were not marked as such were counted as false negatives. See Figure 3.

The inbox contained 33 legitimate e-mails adapted from the Enron e-mail database and 6 phishing e-mails adapted from open sources on recent attacks and our personal inboxes: a malicious Zoom meeting invite and OneDrive file share, two "Nigerian prince"-style scams and two spearphishing examples with urgent requests. Through this setup, we were able to compute participants' overall phishing detection precision (i.e., proportion of true positives out of all e-mails marked as phishing) and recall (i.e., proportion of all phishing e-mails marked as phishing) rates, as well as their detection of specific phishing e-mail types. The total number of

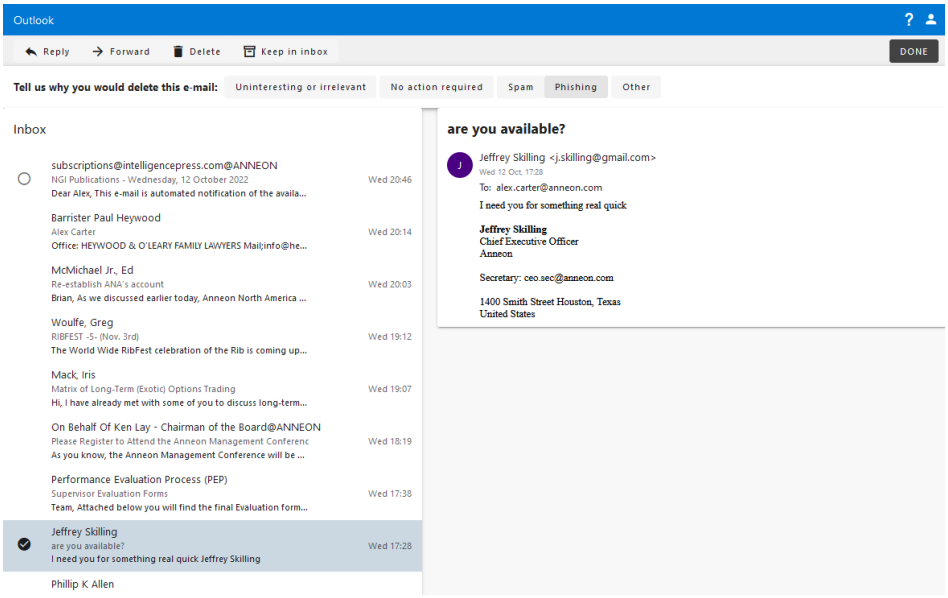


Figure 3: Screenshot of an example spearphishing e-mail in the simulated inbox that is about to be labelled as "Phishing" after clicking "Delete". The interface mimics that of the Outlook web client.

e-mails and phishing proportion was determined to reflect the typical imbalance of phishing versus legitimate e-mails in real inboxes, and to be processed within 30 minutes. Task instructions were always available to participants via the "Help" icon in the top right corner. After processing all e-mails, participants were instructed to click "Done". They then answered demographics questions (age, gender, occupation, income level, industry, organisation size, years of professional working experience) and what they thought was the purpose of the study.

3.4.2 Experiment 2: Simulated phishing attacks. To measure phishing detection in a more ecologically valid way, we worked with the partner organisation's IT Security team to use Office 365 to send simulated phishing attacks to the participants two weeks after they completed the training. We selected five different phishing e-mail types to represent a diverse set of phishing threat models that did not contain obvious layout or language errors: 1. a OneDrive file sharing notification with phishing URL, 2. a job offer with a malicious URL, 3. Office 365 password reset warning, 4. credential harvest through malicious Zoom meeting invitation link, 5. fake conference invite (closest example to spearphishing if recipient's experience happened to relate to the conference topic). After confirming that a participant completed the training task, they were randomly assigned one of the five attack types. At the time of the study, our partner organisation had not sent simulated phishing tests in over six months, and before that phishing simulations were conducted at most once per year. Sending simulated phishing e-mails is part of the IT Security's duties and the phishing e-mails were similar to those used in previous iterations.

The Office 365 phishing attack simulator tracks which users were compromised by, or read, replied to, reported or deleted the simulated attack, where compromised users clicked on a malicious URL, download a malicious attachment, provided credentials on a spoofed web page and/or replied to the e-mail. We found in preliminary testing that the "read" status was unreliable, however, and that participants may delete or ignore an e-mail merely based on the subject line and never opened the full e-mail. As a result, we could not exclude participants based on the read receipts and refrained from using it in the analyses. Hence, we computed the overall compromise rates across all users per attack type and task condition, and performed a Fisher's exact test to infer if any difference in compromise rates between training conditions is statistically significant. Compromised participants were informed about the phishing simulation and they were suggested to review our partner organisation's phishing awareness materials.

4 RESULTS

We performed two experiments to see if an adversarial approach could improve people's ability to detect phishing e-mails. To this end, we developed an adversarial mindset training that engages people with writing phishing e-mails from an attacker's perspective. We then tested if it led to better phishing detection compared to people that received no additional training (Experiment 1) or a conventional phishing awareness training (Experiment 2). The latter was inspired by existing mandatory phishing detection education.

4.1 Training feedback

After watching the respective training video, participants in the conventional (Experiment 2) and adversarial (Experiment 1 and 2) training groups were asked to describe in one word how they felt about the video. These responses are visualised in word clouds in Figures 4a and 4b to give an impression of people's sentiments. The words from the adversarial group lean toward more negative and intense emotions.

(a) Conventional training video (Experiment 2). **(b) Adversarial training video (Experiment 1 and 2).**

Figure 4: Word clouds of the one-word descriptions of how participants felt after watching the training video. Text size corresponds to word frequency.

From: Susan Smith
From e-mail: ssmith@gmail.com

Hi Adam,
Hopefully you remember me from the conference last week. I am in very excited as i have a fabulous business opportunity for you . I am sure a man of your knowledge will be interested. Its concerning a quick payback in the property sector in Manchester. I need a small investment of £20k to secure the final payment on a million pound deal. Due to a brother illness i am not able to ask my family at this time and have reached my limit with the banks. If you are willing to assist me then i am able to offer you a 50% return on your money for just 2 weeks of the loan. Please let me know by return as i need to secure this today.
Kind regards
Susan

Example e-mail 1: Scenario 1: Nigerian prince (advanced fee) scam, response from experiment 1

Overall, participants in the adversarial training group in both Experiment 1 and 2 seemed engaged and empathised with the three phishing writing scenarios. A randomly selected response for each scenario is presented in E-mails 1–3. The full data set is available upon request.

After finishing the adversarial training, participants expressed concerns over whether the written e-mails were convincing enough, whether they would be used in real scams, discomfort when writing the e-mails—although some found it fun to use their creativity—and how easy it is to write such e-mails. All anonymous feedback responses are available via Open Science Framework (OSF).

4.2 Experiment 1 (pilot)

Fifteen participants (mean age = 34.4 (SD = 11.3), 35.7% female) completed the adversarial training and processed e-mails in a simulated inbox two weeks later. Twenty-four participants (mean age = 37.0 (SD = 12.4), 54.2% female) in the control group only performed the e-mail processing task. We computed all participants' phishing detection ability in terms of their precision and recall. Overall, we find that those who received the adversarial training tended to have a better recall than those who received no training ($t(28.1) = -1.48, p = .150$, Cohen's $d = -0.50$, $1 - \beta = 0.32$; Figure 5a), while the overall precision was comparable across groups ($t(26.3) = 1.16, p = .255$, Cohen's $d = 0.40$, $1 - \beta = 0.22$; Figure 5b). That is, participants who received adversarial mindset training detected more phishing e-mails compared to the group that received no training, without becoming significantly less precise in their judgements (e.g. overly labelling e-mails as phishing).

Next, we looked at what phishing e-mail types the adversarial training group was particularly better at detecting compared to the control group. We found that they were better at detecting the two spearphishing e-mails compared to the no training group, both in terms of recall ($t(23.9) = -1.91, p = .069$, Cohen's $d = -0.67$, $1 - \beta = 0.51$; Figure 6a) and precision ($t(22.4) = -2.10, p = .047$, Cohen's $d = -0.75$, $1 - \beta = 0.60$; Figure 6b). Since the adversarial mindset training focused primarily on targeted phishing e-mails (i.e., spearphishing), these findings strongly suggest that engaging

From: Fran Kappali
From e-mail: support.it.un@gmail.com

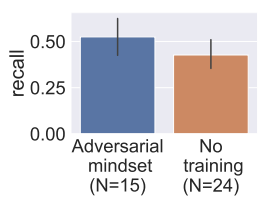
Dear Robyn,
I'm contacting you because we have a report that there has been problems in the UN mail server, some account log in have been failing and your password could be affected by this malfunctions in the server. The whole project management department is currently experience an update in the mail account and password. We have a problem with access to some IT data base, and we expect you cooperation for the matter, we would need your current password and user number so we can verify if it has not been experiencing problems. We hope your soon response and cooperation
IT department specialist
Fran Kappali

Example e-mail 2: Scenario 2: Obtaining login credentials, response from experiment 2

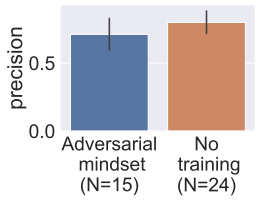
From: James at IBM
From e-mail: J.S.ltd@ibm.com

Hey pete!
Me and some of the guys have managed to se up a quickfire quiz round after lunch to get people hyped up for the rest of the event and thought it would help to throw in some actual prizes too! Could you quickly get together 5 £100 Amazon vouchers for me and send them via email - ill print them out over lunch! I'll pay you back as soon as were back in the office.
thanks again, James

Example e-mail 3: Scenario 3: Get Amazon gift vouchers, response from experiment 1

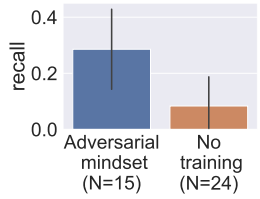


(a) Overall phishing detection recall.

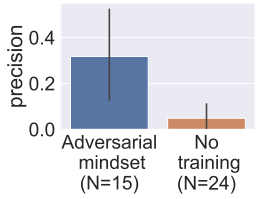


(b) Overall phishing detection precision.

Figure 5: Adversarial mindset training tends to improve overall phishing detection recall ($t(28.1) = -1.48, p = .150$, Cohen's $d = -0.50$, $1 - \beta = 0.32$) compared to having no training.



(a) Spearphishing detection recall.



(b) Spearphishing detection precision.

Figure 6: Adversarial mindset training specifically improves spearphishing detection precision ($t(22.4) = -2.10, p = .047$, Cohen's $d = -0.75$, $1 - \beta = 0.60$) and tends to improve recall ($t(23.9) = -1.91, p = .069$, Cohen's $d = -0.67$, $1 - \beta = 0.51$) compared to having no training.

people with an attacker's perspective helps improve their detection ability.

These results were encouraging enough to pursue Experiment 2 with simulated phishing attacks, especially given the small and imbalanced samples of the pilot study.

4.3 Experiment 2

Given the encouraging results from Experiment 1, we aimed to examine the efficacy of our adversarial mindset training with a more ecologically valid way to measure phishing detection: sending participants simulated phishing e-mails. Furthermore, we compared the efficacy of the adversarial training with a conventional training approach that contained common phishing detection advice and aligned with the organisation's existing training material.

One-hundred-and-forty-four participants (aged between 18 and 45) completed either the adversarial ($N = 40$; 57.5% female; 31.7% in technical department) or conventional ($N = 104$; 73.1% female; 35% in technical department) phishing training and were sent one of five phishing e-mail simulations two weeks afterward. Both groups rated their respective training videos as equally engaging ($t(83.3) = 0.581$, $p = .563$) and useful ($t(91.3) = -0.714$, $p = .477$).

Overall, 15 out of the 104 (14.4%) conventional training group participants and 2 out of the 40 (5%) adversarial training group participants fell for the simulated phishing attacks. That is, people who completed conventional training were nearly three times as likely to fall for a simulated phishing attack compared to those who completed the adversarial training. The difference in the number of people who were compromised between the training groups was at a trending significance level with a small effect size (Fisher's exact test = 0.314, $p = .154$, Cohen's $h = .328$, $1 - \beta = 0.29$; see Table 1). The lower statistical power is mostly due to the imbalanced sample sizes.

Table 1: Number of people (not) compromised per training condition.

EXPERIMENT 2	Not compromised	Compromised
Conventional	89 (85.6%)	15 (14.4%)
Adversarial mindset	38 (95%)	2 (5%)

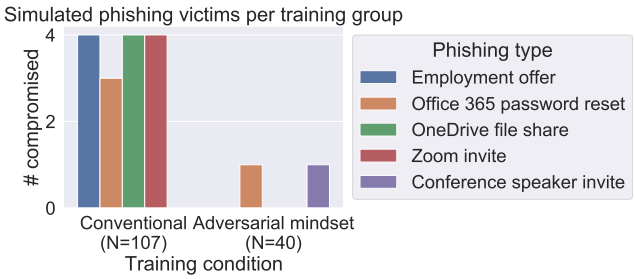


Figure 7: Victimisation count per simulated phishing type for each training condition.

Victims in the conventional training group fell almost equally for all but one of the five simulated phishing types. It is possible that nobody in this group fell for the fifth phishing type, as it purported to invite recipients to speak at a medical conference and most participants did not work in a medical research context. In the adversarial training group, one person fell for it, and one person fell for an e-mail requesting recipients to reset their Office 365 password due to security reasons. See Figure 7.

Three out of the 40 adversarial (7.5%) and nine out of the 104 conventional (8.7%) training participants reported the simulated phishing e-mails. This was not a statistically different proportion (Fisher's exact test = 0.866, $p = 1$, Cohen's $h = .044$). Thus, the training conditions did not affect the amount of people who reported phishing e-mails. That is, both proportions were small, despite the conventional training concluding with the advice that participants should report phishing e-mails to IT. Other participants deleted the e-mail instead of reporting it. Sixteen out of 40 adversarial (40%) and 31 out of the 104 conventional (29.8%) training participants deleted the e-mail, of which one individual deleted it after clicking on the phishing link. This was neither a statistically different proportion (Fisher's exact test = 1.565, $p = .321$. Cohen's $h = .215$).

Lastly, we aimed to control for any potential associations between phishing detection, demographics and other personal factors. To this end, we ran a logistic regression predicting whether participants fell for the simulated phishing attack from their training condition, training video ratings (engagement, usefulness), demographics (age, gender, education level, department), number of

previously completed cybersecurity training and which phishing type they were sent. This only revealed three small to medium sized effects of trending significance. Participants from the adversarial training group ($\beta = -1.38$, $z(128) = -1.66$, $p = .098$, OR = 0.25) and those from a technical department ($\beta = -1.68$, $z(128) = -1.87$, $p = .061$, OR = 0.187) tended to be less compromised, and those who identified as male tended to be more compromised ($\beta = 1.32$, $z(128) = 1.95$, $p = .051$, OR = 3.74).

Together, these experiments show promising results of engaging people with an adversarial mindset to improve their phishing detection ability.

5 DISCUSSION

Phishing detection may seem an adversarial arms race, especially with the advent of generative AI. However, if people understand the principles of online deception tactics, they may understand better why they need to look out for certain message contents and "fake" e-mail addresses. Here, we provide consistent evidence suggesting that engaging people with an adversarial mindset, i.e., thinking of fake e-mail addresses and writing phishing e-mails as if they were an adversary, improves their ability to detect phishing e-mails. Specifically, those who received adversarial training performed better than those who received no (Experiment 1) or conventional (Experiment 2) training.

5.1 Retention of learning

In both experiments, participants' phishing detection abilities were measured two weeks after they completed the provided training. Since we expect training effects to remain high after two weeks, we may regard the present results as an upper-bound estimation of the adversarial training's efficacy versus that of conventional and no training. In this view, the proportional difference in Experiment 2 where the adversarial training group was three times less susceptible than the conventional training group may be seen as quite a dramatic improvement.

Prior works on conventional training efficacy suggest that detection abilities are back to baseline (before training) levels after 5–8 months [42]. Although the results from Experiment 2 suggest that adversarial training leads to a steeper initial enhancement in phishing detection ability compared to conventional training, further studies are needed to measure the adversarial training retention over longer periods of time.

5.2 Dual-use of teaching people how to phish

It is inherently necessary to understand how cybercrime works to defend oneself against it. Cybersecurity awareness training usually describes cybercrime operations and signals to identify them in a passive way. Our adversarial mindset training goes one step further by actively engaging participants to put their passively gained knowledge about phishing into practice. While we did not teach participants new skills about phishing, we may have reduced their inhibitions to write phishing e-mails themselves. This is a classical dual-use scenario, as our method can be used for both defence and offence [43, 47]. There is not much guidance from existing literature on how to deal with dual-use scenarios in teaching [36], whereas more is available on the dual-use of academic research itself [32]. To reduce dual-use risks, the adversarial mindset training did not involve any technical knowledge to teach people how to successfully perform a real phishing attack and we debriefed participants to not pursue phishing in real life. There are a myriad of technical defenses against phishing that successful cybercriminals have to navigate [16].

To further mitigate the risks and to inform our participants about the potential risks of our study, we made participants aware of the writing task after the video, so they could decide if they wanted to continue. Fifteen more people initially started the adversarial training task in Experiment 1, but did not finish it, and we obtained an imbalanced sample in Experiment 2 despite months of sustained recruitment campaigns. We could not verify if participants morally objected to continuing the writing task or if they were not interested in this cognitively more demanding task. Nonetheless, we expect that the potential gains from such adversarial training will weigh up to the fraction of people who may be compelled to pursue phishing in real life, provided all go through the same training and other multi-layered security controls are put in place. Given these precautions, we believe that our adversarial training is ethically acceptable.

5.3 Effort required to implement in organisations

In terms of the practical implementation of the adversarial training, we designed it to take less than 30 minutes on average to complete—comparable to conventional phishing awareness training. The difference is that the adversarial training requires active engagement, whereas conventional training approaches (not serious games) typically use more passive teaching styles [7, 20, 44, 51, 53]. As a concept, we expect an adversarial approach to be equally applicable to helping people in other scam contexts, such as phone scams and phishing through SMS or instant text messages, and highly encourage further studies to test the concept in those domains.

5.4 Limitations

Most of our results were at trending significance levels due to the imbalanced sample sizes in both experiments. Moreover, if our adversarial training groups had been of comparable size to the "no training" and conventional training groups, we likely would have obtained higher statistical power. Yet, especially given the complications in acquiring more participants overall, we are optimistic that the present results provide valuable and encouraging insights for researchers and practitioners who seek different ways to improve phishing detection.

Other limitations of Experiment 2 were the unreliable read receipts, having no convincing spearphishing examples since that required tailoring e-mails to individual contexts, not measuring long-term effects and having no control group due to organisational constraints. We did not perform systematic qualitative analyses of the phishing e-mails participants wrote in the adversarial training groups as it was deemed beyond the scope of the study. We encourage further research in this realm, however, to see if there is a relation between how convincing people's phishing e-mails

are, what persuasion tactics they use (see e.g. [39]), and individual differences in detection ability. Participants' feedback data are available via OSF and written e-mails are available upon request.

6 CONCLUSION

Generative AI is already fuelling the threat of sophisticated phishing attacks. As a result, we need to revise conventional phishing detection training that contains unusable advice and provides limited detection improvement. We proposed the concept of an adversarial phishing training and show over two experiments that it improves detection compared to those who received conventional or no additional training. By not teaching people the technical intricacies of how to perform successful real attacks, but merely engaging their thinking with that of a cybercriminal, we found a nearly three-fold reduction in phishing susceptibility. These findings provide an encouraging new perspective to advance cybersecurity training that may be applied to various contexts beyond phishing.

7 ACKNOWLEDGEMENTS

Sarah Zheng is funded by the UCL Dawes Centre for Future Crime. Ingolf Becker is supported by EPSRC [grant number EP/W032368/1]. We thank all our partner organisation's collaborators from UCL's Information Services Division and their Information Security Group, our actor Bennett Arron, and the anonymous reviewers of the initial manuscript.

REFERENCES

[1] Ralph Adolphs. 2009. The social brain: neural basis of social knowledge. *Annual Review of Psychology*, 60, 693–716. DOI: 10.1146/annurev.psych.60.110707.163514.

[2] Sara Albakry, Kami Vaniea, and Maria K. Wolters. 2020. What is this URL's Destination? Empirical Evaluation of Users' URL Reading. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery. ISBN: 9781450367080. DOI: 10.1145/3313831.3376168.

[3] Aurélien Baillon, Jeroen De Bruin, Aysil Emirmahmutoglu, Evelien Van De Veer, and Bram Van Dijk. 2019. Informing, simulating experience, or both: a field experiment on phishing risks. *PLoS ONE*, 14, 12. DOI: 10.1371/journal.pone.0224216.

[4] Moshe Bar. 2007. The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, 11, 7, 280–289. DOI: 10.1016/j.tics.2007.05.005.

[5] Malak Baslyman and Sonia Chiasson. 2016. Smells phishy? An educational game about online phishing scams. In *APWG Symposium on Electronic Crime Research (eCrime)*. DOI: 10.1109/ecrime.2016.7487946.

[6] Boris C. Bernhardt and Tania Singer. 2012. The neural basis of empathy. *Annual Review of Neuroscience*, 35, 1–23. DOI: 10.1146/annurev-neuro-062111-150536.

[7] Jim Blythe, L. Camp, and Vaibhav Garg. 2011. Targeted risk communication for computer security. In *Proceedings of the 16th international conference on Intelligent user interfaces*, 295–298. DOI: 10.1145/1943403.1943449.

[8] Chris Brook. 2022. Fraud cost americans $5.8 billion in 2021. Retrieved Sept. 21, 2022 from https://digitalguardian.com/blog/fraud-cost-americans-58-billion-2021.

[9] Randy L. Buckner and Daniel C. Carroll. 2007. Self-projection and the brain. *Trends in Cognitive Sciences*, 11, 2, 49–57. DOI: 10.1016/j.tics.2006.11.004.

[10] AJ Burns, M. Johnson, and Deanna Caputo. 2019. Spear phishing in a barrel: insights from a targeted phishing campaign. *Journal of Organizational Computing and Electronic Commerce*, 29, 24–39. DOI: 10.1080/10919392.2019.1552745.

[11] Gamze Canova, Melanie Volkamer, Clemens Bergmann, and Benjamin Berens. 2015. Nophish app evaluation: lab and retention study. In *DSS workshop on usable security*. DOI: 10.14722/usec.2015.23009.

[12] Team R Development Core. 2018. A language and environment for statistical computing. Vienna, Austria, (2018).

[13] Bella M DePaulo. 2004. The many faces of lies. *The social psychology of good and evil*, 303–326.

[14] Anjuli Franz, Verena Zimmermann, Gregor Albrecht, Katrin Hartwig, Christian Reuter, Alexander Benlian, and Joachim Vogt. 2021. SoK: still plenty of phish in the sea — a taxonomy of User-Oriented phishing interventions and avenues for future research. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*. USENIX Association, 339–358. ISBN: 978-1-939133-25-0.

[15] C. J. Gokul, Sankalp Pandit, Sukanya Vaddepalli, Harshal Tupsamudre, Vijayanand Banahatti, and Sachin Lodha. 2018. Phishy - a serious game to train enterprise users on phishing awareness. In *CHI PLAY 2018 - Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*. Association for Computing Machinery, 169–181. ISBN: 9781450359689. DOI: 10.1145/3270316.3273042.

[16] B. B. Gupta, Nalin A. G. Arachchilage, and Kostas E. Psannis. 2018. Defending against phishing attacks: taxonomy of methods, current issues and future directions. *Telecommun syst*, 67, 2, 247–267. DOI: 10.1007/s11235-017-0334-z.

[17] Matthew L. Hale, Rose F. Gamble, and Philip Gamble. 2015. Cyberphishing: a game-based platform for phishing awareness testing. In *48th Hawaii International Conference on System Sciences*, 5260–5269. DOI: 10.1109/hicss.2015.670.

[18] Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120, 11. DOI: 10.1073/pnas.2208839120.

[19] 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, 1–41. DOI: 10.1186/s13673-020-00237-7.

[20] Jurjen Jansen and Paul van Schaik. 2019. The design and evaluation of a theory-based intervention to promote security behaviour against phishing. *International Journal of Human-Computer Studies*, 123, 40–55. DOI: 10.1016/j.ijhcs.2018.10.004.

[21] Matthew L. Jensen, Michael Dinger, Ryan T. Wright, and Jason Bennett Thatcher. 2017. Training to mitigate phishing attacks using mindfulness techniques. *Journal of Management Information Systems*, 34, 2, 597–626. DOI: 10.1080/07421222.2017.1334499.

[22] Rupert Jones. 2021. More than £2.3bn lost in a year as scams surge during pandemic. *The Guardian*.

[23] Christian Keysers and David I. Perrett. 2004. Demystifying social cognition: a hebbian perspective. *Trends in Cognitive Sciences*, 8, 11, 501–507. DOI: 10.1016/j.tics.2004.09.005.

[24] Iacovos Kirlappos and M. Angela Sasse. 2012. Security education against phishing: a modest proposal for a major rethink. *IEEE Security and Privacy*, 10, 2, 24–32. DOI: 10.1109/msp.2011.179.

[25] Ponnurangam Kumaraguru, Yong Rhee, Alessandro Acquisti, Lorrie Cranor, Jason Hong, and Elizabeth Nunge. 2007. Protecting people from phishing: the design and evaluation of an embedded training email system. In *Conference on Human Factors in Computing Systems*, 905–914. DOI: 10.1145/1240624.1240760.

[26] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. 2008. Lessons from a real world evaluation of anti-phishing training. *eCrime Researchers Summit*. ISBN: 9781424429691. DOI: 10.1109/ecrime.2008.4696970.

[27] 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. DOI: 10.1145/1754393.1754396.

[28] Daniele Lain, Kari Kostiainen, and Srdjan Čapkun. 2022. Phishing in organizations: findings from a large-scale and long-term study. In *2022 IEEE Symposium on Security and Privacy (S&P)*. IEEE, 842–859. DOI: 10.1109/SP46214.2022.9833766.

[29] Harjinder Singh Lallie, Lynsay A Shepherd, Jason RC Nurse, Arnau Erola, Gregory Epiphaniou, Carsten Maple, and Xavier Bellekens. 2021. Cyber security in the age of covid-19: a timeline and analysis of cyber-crime and cyber-attacks during the pandemic. *Computers & Security*, 105, 102248. DOI: 10.1016/j.cose.2021.102248.

[30] Claus Lamm, Jean Decety, and Tania Singer. 2011. Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *NeuroImage*, 54, 3, 2492–2502. DOI: 10.1016/j.neuroimage.2010.10.014.

[31] Raymond A. Mar. 2011. The neural bases of social cognition and story comprehension. *Annual Review of Psychology*, 62, 103–134. DOI: 10.1146/annurev-psych-120709-145406.

[32] Seumas Miller and Michael J. Selgelid. 2007. Ethical and Philosophical Consideration of the Dual-use Dilemma in the Biological Sciences. *Science and Engineering Ethics*, 13, 4, 523–580. DOI: 10.1007/s11948-007-9043-4.

[33] Mattia Mossano, Kami Vaniea, Lukas Aldag, Reyhan Düzgün, Peter Mayer, and Melanie Volkamer. 2020. Analysis of publicly available anti-phishing webpages: contradicting information, lack of concrete advice and very narrow attack vector. In *IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, 130–139. DOI: 10.1109/EuroSPW51379.2020.00026.

[34] Cybersecurity National Cyber Security Centre (NCSC) and Infrastructure Security Agency (CISA). 2020. Advisory: covid-19 exploited by malicious cyber actors. Retrieved Mar. 26, 2022 from https://www.ncsc.gov.uk/news/covid-19-exploited-by-cyber-actors-advisory.

[35] Sophie J. Nightingale and Hany Farid. 2022. Ai-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119, 8. DOI: 10.1073/pnas.2120481119.

[36] Martin S. Olivier. 2016. On the Morality of Teaching Students IT Crime Skills: Extended Abstract. In *Proceedings of the Computer Science Education Research Conference 2016*. Association for Computing Machinery, New York, NY, USA, 2–3. ISBN: 978-1-4503-4492-0. DOI: 10.1145/2998551.2998553.

[37] Florian Quinkert, Martin Degeling, Jim Blythe, and Thorsten Holz. 2020. Be the phisher – understanding users' perception of malicious domains. In *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*. Association for Computing Machinery, 263–276. ISBN: 9781450367509. DOI: 10.1145/3320269.3384765.

[38] Emilee Rader and Anjali Munasinghe. 2019. "wait, do i know this person?": understanding misdirected email. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–13. ISBN: 9781450359702. DOI: 10.1145/3290605.3300520.

[39] Prashanth Rajivan and Cleotilde Gonzalez. 2018. Creative persuasion: a study on adversarial behaviors and strategies in phishing attacks. *Frontiers in Psychology*, 9. DOI: 10.3389/fpsyg.2018.00135.

[40] Hannes Rakoczy. 2022. Foundations of theory of mind and its development in early childhood. *Nature Reviews Psychology*, 1, 4, 223–235.

[41] Elissa M. Redmiles, Noel Warford, Amritha Jayanti, Aravind Koneru, Sean Kross, Miraida Morales, Rock Stevens, and Michelle L. Mazurek. 2020. A comprehensive quality evaluation of security and privacy advice on the web. In *Proceedings of the 29th USENIX Security Symposium*, 89–108. ISBN: 9781939133175.

[42] 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. *Proceedings of the 16th Symposium on Usable Privacy and Security (SOUPS 2020)*, 259–284. ISBN: 9781939133168.

[43] Thea Riebe and Christian Reuter. 2019. Dual-Use and Dilemmas for Cybersecurity, Peace and Technology Assessment. In *Information Technology for Peace and Security: IT Applications and Infrastructures in Conflicts, Crises, War, and Peace*. Christian Reuter, (Ed.) Springer Fachmedien, Wiesbaden, 165–183. ISBN: 978-3-658-25652-4. DOI: 10.1007/978-3-658-25652-4_8.

[44] Sebastian W. Schuetz, Paul Benjamin Lowry, Daniel A. Pienta, and Jason Bennett Thatcher. 2020. The effectiveness of abstract versus concrete fear appeals in information security. *Journal of Management Information Systems*, 37, 3, 723–757. DOI: 10.1080/07421222.2020.1790187.

[45] Anastasia Shuster and Dino Levy. 2020. Contribution of self- and other-regarding motives to (dis)honesty. *Scientific reports*, 10, 15844. DOI: 10.1038/s41598-020-72255-5.

[46] Mario Silic and Paul Benjamin Lowry. 2020. Using design-science based gamification to improve organizational security training and compliance. *Journal of Management Information Systems*, 37, 1, 129–161. DOI: 10.1080/07421222.2019.1705512.

[47] Julia Slupska and Leonie Maria Tanczer. 2021. Threat Modeling Intimate Partner Violence: Tech Abuse as a Cybersecurity Challenge in the Internet of Things. In *The Emerald International Handbook of Technology-Facilitated Violence and Abuse*. Jane Bailey, Asher Flynn, and Nicola Henry, (Eds.) Emerald Publishing Limited, 663–688. ISBN: 978-1-83982-849-2. DOI: 10.1108/978-1-83982-848-520211049.

[48] Mark Sweney. 2023. Darktrace warns of rise in ai-enhanced scams since chatgpt release. *The Guardian*.

[49] Melanie Volkamer, Martina Angela Sasse, and Franziska Boehm. 2020. Analysing simulated phishing campaigns for staff. In *European Symposium on Research in Computer Security (ESORICS 2022)*. Springer, Cham, 312–328. DOI: 10.1007/978-3-030-66504-3_19.

[50] Rick Wash. 2020. How experts detect phishing scam emails. In *Proceedings of the ACM on Human-Computer Interaction* number 2. Vol. 4. Association for Computing Machinery. DOI: 10.1145/3415231.

[51] 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*. Association for Computing Machinery. ISBN: 9781450356206. DOI: 10.1145/3173574.3174066.

[52] 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*. Association for Computing Machinery. ISBN: 9781450359702. DOI: 10.1145/3290605.3300338.

[53] 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* (HoTSoS). Association for Computing Machinery, Hanover, MD, USA, 52–61. ISBN: 9781450352741. DOI: 10.1145/3055305.3055310.

[54] Sarah Y. Zheng and Ingolf Becker. 2023. Checking, nudging or scoring? evaluating e-mail user security tools. In *Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023)*. USENIX Association, Anaheim, CA, 57–76. ISBN: 978-1-939133-36-6.

[55] Sarah Y. Zheng and Ingolf Becker. 2022. Presenting suspicious details in User-Facing e-mail headers does not improve phishing detection. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*. USENIX Association, Boston, MA, 253–271. ISBN: 978-1-939133-30-4.