Computers & Security 137 (2024) 103620

Contents lists available at ScienceDirect

Computers & Security

journal homepage: www.elsevier.com/locate/cose

Taking 5 minutes protects you for 5 months: Evaluating an anti-phishing awareness video

Benjamin M. Berens *, Mattia Mossano, Melanie Volkamer

SECUSO, Karlsruhe Institute of Technology, Kaiserstraße 12, Karlsruhe, 76131, Baden-Württemberg, Germany

ARTICLE INFO

Keywords:
Anti-phishing awareness
Video measure
Retention study
User study
Phishing knowledge

ABSTRACT

Phishing is one of the biggest security threats to organizations. Anti-phishing awareness measures can improve phishing email detection rates. These measures need to be efficient, effective, and have an enduring impact over months, rather than days. Related research provides evidence of their effectiveness in the short term. However, questions remain as to how long this impact endures. We conducted a retention user study in two phases, with almost 200 participants in the first phase and almost 80 in the second phase, to determine whether a five-minute video retains its effectiveness five months after the intervention (similar to related work on more time-intensive measures). Our results suggest that short videos can indeed still exert a positive influence five months later. We also report on the video's influence on phishing detection strategies, as well as on viewers' confidence in this respect. Based on our results, we propose recommendations to inform the content of future awareness refreshment measures.

1. Introduction

In the third quarter of 2022, Anti-Phishing Working Group (2022) registered the highest phishing attack number since they started collecting data. To address the ever-increasing phishing threat, organizations employ various anti-phishing awareness measures, i.e., interventions to increase awareness of the phishing threat itself and building the skills to resist the deceptive attempt. Various measures are in place, such as instructor-based courses, e-learning, games, and videos – with several proposals coming from the research community, e.g., Chang and Coppel (2020); Reinheimer et al. (2020); Tschakert and Ngamsuriyaroj (2019). However, because the time spent on anti-phishing awareness measures reduces employees' productive hours, these measures should be as efficient as possible while still being effective, i.e., enhance employees' ability to detect phishing emails. The measure should raise skills right after consumption and also persist for several months. Especially as the time spent on renewing employees' anti-phishing knowledge also reduces productive working hours and should therefore be reduced to the necessary minimum. Anti-phishing awareness videos (from now on called "videos") are usually less time-intensive than instructor-based courses, e-learning, and games, while still being effective, as shown for specific videos in, e.g., Abawajy (2014) and Hamdani and Mustafa (2021).

Previous research showed that the participants' phishing detection ability was retained for five months. For example, Reinheimer et al. (2020) showed so for instructor-based course, Canova et al. (2015b) showed it for games, while Berens et al. (2022) showed it for e-learning. Further details on this five-month retention period are in section 2.

In Volkamer et al. (2018), we developed and evaluated a 5-minute video. Given its short duration, it addressed time-investments concerns. Yet, a short measure based on passive knowledge transfer could be less effective than more interactive solutions like courses or e-learning. Our evaluation, though, showed that our video significantly enhanced the participants' ability to detect phishing emails. However, since then we received feedback on the original video and updated it. Furthermore, in Volkamer et al. (2018) we did not consider a five month interval for knowledge retention. Moreover, the phishing landscape is in continuous evolution, and defense mechanisms need continuous updates too. Hence, reevaluating our video also allow us to verify that it can still help to significantly increase the phishing detection ability of viewers.

Our contributions in this paper are:

1. Improving the video from Volkamer et al. (2018) based on the feedback received from participants of the original study, as well as from security experts, to further improve its effectiveness with respect to the ability to detect phishing emails.

* Corresponding author.
E-mail address: benjamin.berens@kit.edu (B.M. Berens).

https://doi.org/10.1016/j.cose.2023.103620
Received 17 August 2023; Received in revised form 15 November 2023; Accepted 20 November 2023
Available online 29 November 2023

(a) Example of a tooltip.

(b) Example of a status-bar.

Fig. 1. Visual representation of the two forms to display the URL (status-bar and tooltip).

2. Evaluating the effectiveness of the improved version of the video with respect to the ability of 79 participants to detect phishing emails.
3. Determining that the increased phishing detection ability is still statistically significant after five months.
4. Gaining insights on the impact of the video on the strategies used by viewers to detect phishing emails.
5. Gaining insights on the impact of the video on viewers' confidence in their ability to detect phishing emails.
6. Gaining insights into the aspects that should be addressed in potential refreshment measures.

Using an improved version of the video, we designed and conducted a within- and between-subjects study with 79 German participants recruited from the Clickworker panel service. Our focus was on emails containing at least one link which was also the only indicator of a phishing attack. The participants were distributed in four groups, depending on whether they watched our video or not, and whether they judged the emails (see Fig. 1 for visual representations of a tooltip and a status-bar) in a ***tooltip*** & the status-bar interface (similar to, e.g., Microsoft Outlook) or a ***status-bar*** interface (similar to, e.g., web browsers or Mozilla Thunderbird). After five months, we recalled our participants and asked them to judge interactive screenshots again. Note, the URL was only displayed on mouse hover of the link. Also note that, for readability, we will use "tooltip" instead of "tooltip & status-bar" in the remainder of this paper.

This study design led us to interesting answers of the goals listed above. Regarding the second goal, we discovered that the video improved the phishing detection of participants independently of the interface used. However, after five months (third goal), only participants in the tooltip condition were still significantly better, suggesting that it could be better for security awareness retention to use tooltips, instead of other interfaces. It was also interesting that the video seemed to impact the detection strategies reported by participants right after watching it. Yet, this was no longer the case after five months, although the effect was still detectable on their performance (fourth goal). Regarding the fifth goal, watching or not the video seemed to have no effect on the participants' certainty, who were in general overly confident in their abilities. Finally, it seemed that if any refreshment measure was to be provided, it should focus on reminding users that even slight variations in URLs are important, e.g. `arnazon.de` instead of `amazon.de`.

In conclusion, we determined that a 5-minute video might be sufficient to significantly increase the ability to detect phishing emails, as we measured a significant retention level similar to that of far more time-intensive (and extensive) measures proposed and evaluated in related work. We also provide some recommendations on refreshment measures based on the study results.

2. Related work

Phishing awareness measures in general Phishing awareness measures come in different formats: games, on-site instructor-based tutorials, e-learning, texts of various lengths, and videos. Table 1 offers some examples of different formats from literature. An overview of anti-phishing awareness measures is provided by Franz et al. (2021) and Jampen et al. (2020). In particular, we first tried to incorporate in our work how to conduct an evaluation of a phishing awareness measure, based on some of the related work: Berens et al. (2022), Canova et al. (2015a,b), Neumann et al. (2017), and Volkamer et al. (2016). We then combined it with findings from the literature about knowledge retention to achieve our research goals (see the section on Retention of anti-phishing awareness later on).

Table 1
A short overview of type of phishing awareness measures in literature. Video measures are highlighted with a black diamond.

Authors	Type of measure	
Abawajy (2014)	Video	♦
Althobaiti et al. (2021)	Textual	
Arachchilage et al. (2014)	Game	
Arachchilage et al. (2016)	Game	
Berens et al. (2022)	E-learning	
Canova et al. (2015a)	Game	
Canova et al. (2015b)	Game	
Gokul et al. (2018)	Game	
Chang and Coppel (2020)	Instructor based	
Garg et al. (2011)	Video	♦
Gonzalez and Locasto (2015)	Textual	
Hamdani and Mustafa (2021)	Video	♦
Hart et al. (2020)	Game	
Kunz et al. (2016)	Game	
Lastdrager et al. (2017)	Textual	
Misra et al. (2017)	Game	
Neumann et al. (2017)	Textual	
Onashoga et al. (2019)	Game	
Reinheimer et al. (2020)	Instructor based	
Sheng et al. (2010)	Textual	
Tschakert and Ngamsuriyaroj (2019)	Instructor based	
Volkamer et al. (2016)	Textual	
Wash and Cooper (2018)	Textual	
Wen et al. (2019)	Game	
Zielinska et al. (2014)	Textual	

Advantages and disadvantages of videos There are two main reasons to focus on videos: (1) videos can be very short while still covering a lot of content, also allowing the use of animation and visualization of concepts. (2) Hamdani and Mustafa (2021) showed that video measures increase viewers' engagement and attention, as compared to text-based measures. Abawajy (2014) also showed that both games and video measures have a higher effectiveness than text-based measures. Still videos as a passive measure lack some features that both the games and more extensive training provide: 1) they can give feedback to the user performance and 2) they can include exercises where participants can test themselves for remaining gaps.

Evaluated anti-phishing awareness videos Various videos exist (e.g., when searching on YouTube), but only a few have been evaluated.

Table 2
Related work on phishing detection rate after certain time intervals. We only report those time intervals where the rate is still significantly better than the baseline, i.e., before and after the measure.

Paper	*Volkamer et al. (2018)*	*Reinheimer et al. (2020)*	*Canova et al. (2015b)*	*Berens et al. (2022)*
Measure type	Video	With Instructor	Game	E-learning
Duration (minutes)	5	180-240	30	108
Time interval (months)	2	4	5	5
Baseline (Mean)	42.60%	62.00%	57.24%	63.24%
Post (Mean)	86.90%	80.00%	90.79%	90.44%
Diff. Baseline-Post	+44.30%	+18.00%	+33.55%	+27.20%
Retention (Mean)	81.30%	71.00%	81.89%	84.93%
Diff. Baseline-Retention	+38.70%	+9.00%	+24.65%	+21.69%
Diff. Post-Retention	-5.60%	-9.00%	-8.90%	-5.86%

For example, Garg et al. (2011) developed a video aimed at raising awareness in older adults and evaluated it with 12 participants. Abawajy (2014) compared various anti-phishing awareness measures (one text, one video, and three games) to determine which delivery method was preferred. Neither the video from Garg et al. (2011) nor the one from Abawajy (2014) is available, and, as such, we could not base our research on them. Hamdani and Mustafa (2021) compared the effectiveness of different anti-phishing awareness delivery measures (a website, three videos and an info-graphic) in a user study with 78 participants. Of these three videos, only two are still available, but have several issues. The first[1] is 2:53 minutes long, it shows a phishing mail and highlights the phishing cues in the email while providing information in a voice-over. However, it also shows the action of clicking on the embedded link to visit the phishing website and uses it to highlight the different phishing cues in the email. Showing the clicking action is sub-optimal, as it might give the impression that it is safe to click on a link as long as one checks the web address. In reality, the simple action of clicking on a malicious link could trigger a direct-download attack, i.e., malware being installed on the victim's device, as explained, e.g., in Singhal and Levine (2019) and in Sood and Zeadally (2016). Hence, we decided not to use this video. The second video[2] is 4:54 minutes long and presents a series of recommendations on how to detect phishing emails, with examples and audio descriptions. Although it is well executed, it does not highlight the location of the status-bar and does not consider phishing URLs with the correct domain in the path or sub-domain (as opposed to the domain). It is assumed that the viewers are capable of parsing URLs, which is not the case, as shown by Albakry et al. (2020). We base our research on the video proposed in Volkamer et al. (2018), which is 5:04 minutes long and it proved effective in enhancing the phishing detection of users, without suffering from any the above-mentioned limitations.

Retention of anti-phishing awareness Some studies evaluated the impact of their phishing awareness measures on the phishing detection rate not only straight after delivery, but also after a delay. Reinheimer et al. (2020) conducted a study in a German public enterprise employing an on-site tutorial. The tutorial addresses three topics: (1) general security awareness, (2) phishing, and (3) password 'best practice'. This is considerably more time-intensive (around three to four hours) than the other measures mentioned in this paragraph. The authors report that the effect was still significant after four months. In Canova et al. (2015b), the focus is on a game. The authors conducted a retention study after 5 months and found that the app effect was still significant. Berens et al. (2022) employed e-learning in a German University. The authors report that the effect was still significant after five months. In Volkamer et al. (2018), we reported on a study of a 5 minute video and showed that its effect was still significant after 2 months. However, the results from related work (see in particular Table 2) indicate that anti-phishing awareness measures can have a significant effect at least up to five months. Hence, we wanted to evaluate how reliably a short awareness measure would enhance phish detection after five months, especially when compared to the more time-intensive measures, such as those reported by Berens et al. (2022), Canova et al. (2015b), and Reinheimer et al. (2020).

Besides the fact that the original video was not evaluated after a five month delay, there are other reasons for conducting the research presented in this paper: (a) we received valuable feedback from both participants and security experts to improve the video (more on this in the next section); (b) the participants were young (all below 36); (c) the evaluation conducted in Volkamer et al. (2018) was based on screenshots which always displayed the URL behind links. A more realistic evaluation would require participants to hover over the link themselves to see the URL.

It is worth mentioning the retention study carried out by Kumaraguru et al. (2009). The authors evaluated the knowledge retention elicited by an embedded training system, PhishGuru, that sends simulated phishing emails to users. If a user clicks on the embedded link, they are taken to a page that delivers an infographic as an anti-phishing measure. However, if the user does not click on the embedded link sent by PhishGuru, they will not see the anti-phishing awareness measure. For this reason, it might be that some participants are never exposed to the anti-phishing awareness measure. Moreover, they only check retention after approximately a month, which is already covered by the other related research we mentioned. For these reasons, we acknowledge their work, but do not extend it.

In summary, a current study on such a short and easy to distribute measure as the video is currently missing. There is also a lack of results on how such a short and passive measure performs over a longer period of time without any form of exercises.

3. Research questions

Our overarching **research goal** was to evaluate the improved video with regard to the ability to detect phishing mails five months after viewing the video. To achieve this, we considered two different ways of displaying the URL behind a link: status-bar and tooltip. We considered both because: (1) both exist in the real world (i.e., status-bar for the Thunderbird and well-known web browsers; and tooltip in Outlook), and (2) there might have been significant differences between the performances due to the tooltip being closer than the status bar to the user's focus when clicking a link.

Our research goal was split into seven **sub-goals**, each mapped to one or more research questions (hereafter, *RQ*):

1. Show that the participants' ability to detect phishing emails significantly improved immediately after watching the new version of our 5 minutes video – i.e., confirming the results from our previous study from Volkamer et al. (2018)) (RQ1 and RQ2);
2. Determine whether the hypothesized enhanced ability to detect phishing emails was still evident after five months - the time interval according to related research (RQ3 and RQ4);

[1] https://www.youtube.com/watch?v=fyfAKQM3qTY - Last checked: 09.11.2023.
[2] https://www.youtube.com/watch?v=U7tbJVSInvo - Last checked: 17.08.2023.

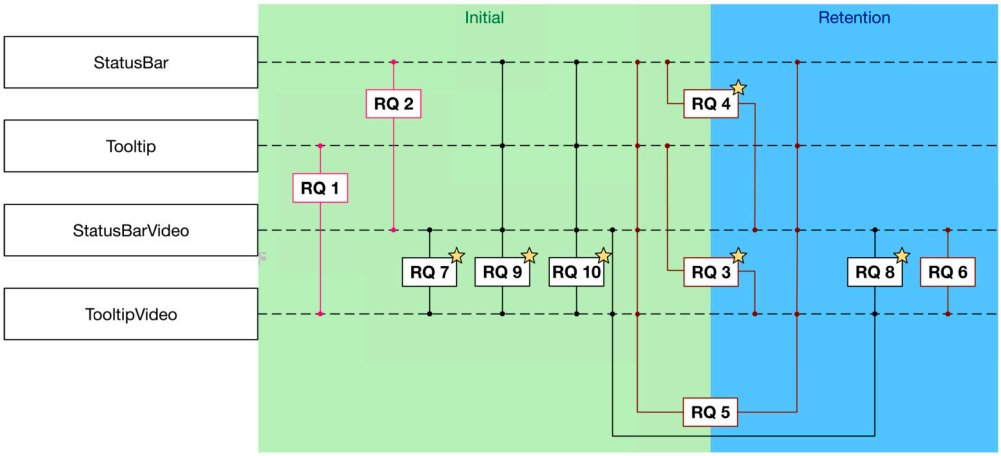

Fig. 2. Overview of all research questions with links between groups and phases. The main questions are colored in red, while the prerequisite ones are in pink. The further findings are in black. Stars mark those research questions that are our main contributions.

3. Identify factors users considered in making their decision directly after watching the video and at retention phase (RQ5);
4. Identify areas where a refreshment is needed after five months to improve the phishing detection (RQ6);
5. Determine whether the impact of our video was different depending on the interface seen (RQ7 and RQ8);
6. Explore whether our video changed participants' confidence level (RQ9);
7. Find if there is a difference between phish detection performance in the presence of Status-bar or Tooltip (RQ10);

Fig. 2 shows an overview of when the RQs were investigated over the two phases, and which study groups were involved (discussed in section 3.4).

3.1. Pre-requisite research questions

Our first two RQs aim at confirming the results of Volkamer et al. (2018), and they are effectively pre-requisites for the other RQs. This because, if the video has no effect right after watching it, then no effect can be expected after five months.

RQ1: TooltipVideo effectiveness When the URL is displayed in a tooltip & status bar interface, is the ability to detect phishing emails of users significantly higher after watching the video?

Volkamer et al. (2018) showed that the participants who watched the video exhibited an improved ability to detect phishing emails right after watching it. Hence, we phrase the hypothesis $\mathbf{H}_1$: Participants using the tooltip interface that watched the video have a significantly better ability to detect phishing emails than participants using the tooltip interface that did not watch the video.

RQ2: StatusBarVideo effectiveness When the URL is displayed in a status-bar interface, is the ability to detect phishing emails of users significantly higher after watching the video?

Just as in the previous RQ, we build on the results of Volkamer et al. (2018) to phrase hypothesis $\mathbf{H}_2$: Participants using the status-bar interface that watched the video have a significantly better ability to detect phishing emails than participants using the status-bar interface that did not watch the video.

3.2. Main research questions

In this section we present our main RQs. These are the focus of our study.

RQ3: TooltipVideo effectiveness after 5 months (retention) When the URL is displayed in a tooltip, is the ability to detect phishing emails of users still significantly more effective *5 months* after watching the video?

Our hypothesis here was $\mathbf{H}_3$: Participants using the tooltip interface that watched the video still have a significantly better ability to detect phishing emails after 5 months than participants using the tooltipinterface without watching the video. This hypothesis was based on the findings from related work (see section 2) indicating that the phishing detection rate was still significant after 5 months. Furthermore, in our previous study, we showed that the phishing detection rate of participants that watched the first version of our video was still significant two months after watching the video (at the initial phase).

RQ4: StatusBarVideo effectiveness after 5 months (retention) When the URL is displayed in a status-bar interface, is the ability to detect phishing emails of users still significantly higher *5 months* after watching the video?

Similarly to RQ3 (and with the same justification), we formulated hypothesis $\mathbf{H}_4$: Participants using the status-bar interface that watched the video still have a significantly improved ability to detect phishing emails after 5 months than participants using the status-bar interface without watching the video (at the initial phase).

RQ5: How decided What do users consider when deciding whether an email is a phish or not - without the video, with the video, and 5 months after watching the video?

This is a purely exploratory question, hence we formulate no hypothesis. To answer this RQ we asked our participants how they judged the email screenshots in an open text question, and we then analyzed their responses with open coding (see section 5.5.3).

RQ6: Refreshments Is it required to contain all the phishing techniques in the refreshment measures or can time be saved by focusing can on those with lower performance?

Just as RQ4, we formulate no hypothesis in this case either. Rather, we answered the RQ by checking the different phishing tricks descriptive statistics performance to give more concrete points to address for future refreshments.

3.3. Further research questions

This section presents RQs that are related to the main ones and that we were also interested in answering, albeit they are not the main focus of our work.

Except for RQ10, we did not formulate hypotheses for any of the other RQs in this section. This is because we had no strong basis to expect a certain outcome instead of another.

RQ7: Tooltip vs. StatusBar - influence of video Is the effect of the video on the ability to detect phishing emails of users of the status-bar inter-

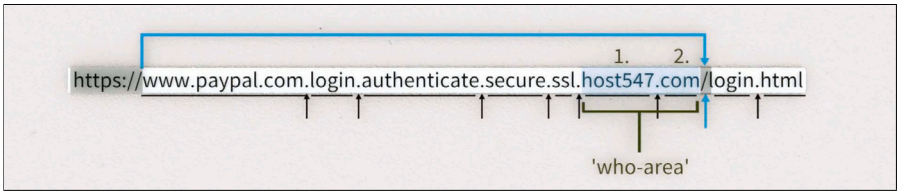

Fig. 3. URL with highlighted 'who-area'.

face different from users of the tooltip interface - right after watching it?

RQ8: Tooltip vs. StatusBar - influence of video (retention) Is the effect of the video on the ability to detect phishing emails of users of the status-bar interface different from the one of users of the tooltip interface - five months after watching it?

RQ9: Certainty - influence of video Is the certainty of users of the tooltip interface different than the certainty of users of the status-bar interface? Is the certainty of correct answers different from that of wrong answers?

Confidence is becoming increasingly important in cyber security awareness. Votipka et al. (2020) showed that a part of confidence is the belief in their ability to perform a task. Additionally our task was a binary decision between phishing or legitimate. Therefore, we asked our participants how certain they were regarding their decision for every example, so to get more insight into their confidence in their ability to detect a phish. We also wanted to see if wrong decisions were due to uncertainty and whether right decisions were more leaning towards certainty. We expected the video to positively influence the participants' certainty, as they would have advice they could base their decision on provided merely a few minutes ago. For the actual difference between right or wrong decisions we wanted to exploratory investigate if there are differences in the first place and, if so, whether the video has an additional influence on the different kind of decisions.

RQ10: Tooltip vs. StatusBar - general performance Is the ability to detect phishing emails of users of tooltip interface better than the one of users of the status-bar interface?

In this case, our hypothesis was $\mathbf{H}_{10}$: Participants using the tooltip interface have a significantly better ability to detect phishing emails than participants using the status-bar interface. We hypothesized this because Petelka et al. (2019) and Volkamer et al. (2016) showed that placing the URL just-in-place (i.e., in a tooltip next to the link to analyze) is effective in increasing the users' ability to detect phishing emails. This is not true for a status-bar.

3.4. Study groups

We considered the following study groups:

StatusBar Participants in this group saw the URL in a status-bar in the bottom left-hand corner of the screen when hovering over a link with their mouse cursor.

StatusBarVideo Similar to the *StatusBar* group, except that participants watched the video described in section 4.1 before judging the emails.

Tooltip Participants in this group saw the URL in a simple tooltip, appearing next to the link when hovering a link with their mouse cursor, and in a status-bar similar to the one shown in the *StatusBar* group. In addition to the URL, the tooltip displayed the sentence: "Click or tap to follow the link". As such, it was similar to the one used in the MS Outlook client.

TooltipVideo Similar to the *Tooltip* group, except that participants watched the video described in section 4.1 before judging the emails.

4. Methodology

In this section we introduce the video that we employed and the improvement it underwent after its first evaluation in Volkamer et al. (2018), and we describe the design of our study, including recruitment and ethical considerations. Note that the study methodology is described in English with international examples (e.g., `amazon.com` instead of `amazon.de`). Yet, the study was conducted in German with German participants and country-specific emails and URLs.

4.1. Anti-phishing awareness video

In Volkamer et al. (2018), we report on the development and evaluation of a video that did not merely raise awareness of phishing threats but also explained how to detect phishing emails. The focus is on phishing with embedded links redirecting users to the phishers' web server. The intended audience is the general public (hence, no sophisticated technical knowledge or familiarity with terminology is assumed).

The video first raised awareness for two types of phishing attacks: (1) Phishing where attackers deceive users through authentic-looking phishing emails with embedded links in order to *either* (a) access sensitive information once entered on the web page behind the link *or* (b) to spread malware once victims clicked on it. (2) It explained that the URL behind links should be checked before clicking and that the URL only appeared either in the status-bar or in a tooltip once one hovers over the link. After that, two tips are introduced on how to check URLs: *Tip 1* explained the importance of focusing on the domain&top level domain[3] which is called "who-area" in the video (see Fig. 3). *Tip 2* explained the importance of checking the who-area letter by letter to detect URLs such as `paypal.de`. Then, an example showed that a link can look like a URL yet be different from the actual URL behind this link.

The video was evaluated in a between-subjects online study with 89 participants. Participants were asked to judge emails either as phishing or legitimate before watching the video, straight after watching the video and eight weeks later. The ability to detect phishing emails significantly increased right after the video. For those who also participated after eight weeks (i.e., almost 2 months), the detection rate was still significantly higher than before watching the video. However, the impact of the video was not evaluated after a five month interval to test for knowledge retention.

As part of this first study, we collected feedback to help us to improve the video. Afterwards, we also showed the video to security experts who have published papers on phishing awareness and asked their feedback. This was done informally with a request to focus on correctness and completeness. Based on the feedback, we implemented several changes:

- Using existing services such as PayPal and DHL as illustrations in the video could potentially lead to legal issues, particularly when making it publicly available. Using these in the emails is still acceptable, since they are not being published. Therefore, the reworked video used fictional services.

[3] Note, as the video was developed for Germany this would be `example.de`; in the U.K. this would need to be changed to `example.co.uk`.

(a) Original video (b) Reworked video

Fig. 4. Coloring of the URLs in the two video versions.

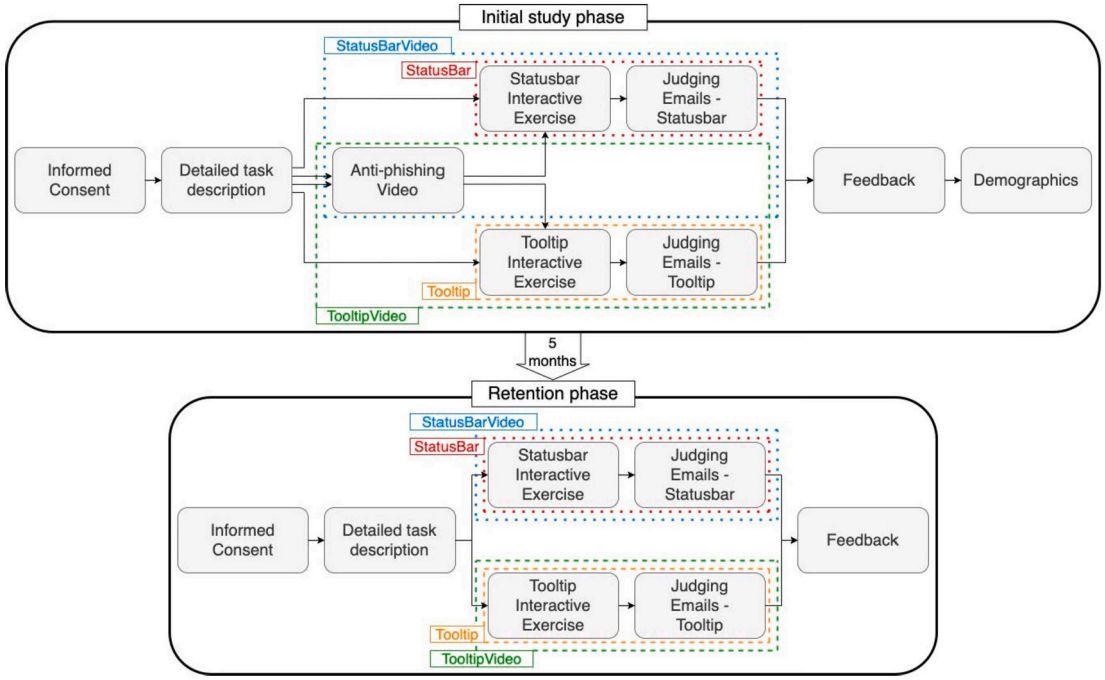

Fig. 5. Flowchart of the study design with the participants groups.

- We replaced some example URLs used to explain Tip 1. The original video focused on phishing URLs where the correct domain was part of the subdomain. This was changed to put further emphasis on phishing URLs where the correct domain appears as part of the path. In addition, the reworked video extends the display time of these examples by three seconds.
- Another change concerned the highlighting of different URL parts related to a URL who-area identification. In the original video, the who-area was highlighted in green. The experts advised against this color scheme as highlighting a fake domain in a color generally perceived as positive could be misleading. The reworked video used blue as a neutral alternative color. Fig. 4 shows the corresponding changes.
- The experts also made us aware that the URL `linkedin.com`, used as example of URL manipulation, is registered by someone other than LinkedIn. To prevent the risk of someone checking the URL, it was replaced with `0stermarkt.de` in the reworked video.
- Later in the video, different short URLs were shown to the viewer. As some of those services have updated their pages to use `https` instead of `http`, we updated the respective examples as well.
- In the final scene, the video showed a link to further information on this topic. This link text was a URL. The mouse hovers over this and the real URL behind the link were displayed. The real URL was different from the URL in the link text. It was explained that it is important not to trust the URL in the link text. In the original video, the actual URL was displayed in a tooltip. The experts suggested replacing the tooltip with a status-bar, as in such contexts it would be more difficult to notice the mismatch between the URL in the link text and the actual URL. For the reworked video, this scene was changed in response to the feedback.
- The last change was the addition of a summary at the end of the reworked video. It repeated the different phishing techniques shown throughout the video and was added based on the feedback we received.

The new version of the video was still only 5:03 minutes long. The video is freely available, both in English[4] and in German.[5]

4.2. Survey design

Our study was both a within- and between-subjects study, with participants randomly assigned to one of four groups (see section 3.4). We used the online survey platform SoSci Survey[6] to collect the data, because the company is compliant with the European Data Protection Regulation (GDPR).

The study was split into two phases: initial phase and retention phase. An overview of the two phases is provided in Fig. 5 and the details described in the following sections.

4.2.1. Initial study phase

The first part of our study consisted of a survey that every participant completed. Participants were told right before they started that the survey included attention questions and that they needed to be completed to finish the study. The steps of this initial phase are:

Informed consent Participants were informed about the general goal of the study, that there was an exclusion exercise further on, and what their rights were. This included the fact that the data would be anonymously analyzed. They were also informed that they could stop the

[4] https://www.youtube.com/watch?v=1phRPBIjFoo - Last checked: 17.08.2023.
[5] https://www.youtube.com/watch?v=JYu07OcFzew - Last checked: 17.08.2023.
[6] https://www.soscisurvey.de/en/index.

study at any time, without providing a reason, in which case the data would not be used. The participants were then asked to give their consent before proceeding. The translated informed consent document is available in the supplementary material.

Detailed task description In the second step, the participants received a detailed task description. The specific texts were different depending on the group the participant was assigned to: the two video groups were informed that they would watch a video about phishing and that this video was important later on in the study. They were asked to turn on the sound. Furthermore, they were told that they would be judging emails in the Chrome browser, seen in the Gmail account of Martin Müller. The interface they used depended on their group, i.e., status-bar or tooltip; participants were not informed of the differences. All groups, except for the status-bar, were told that the interface was a browser update. The groups without the video only received information regarding the email judgment. Participants were told that all services mentioned in the email screenshots were familiar to Martin Müller (i.e. he actually has accounts there).

All participants were informed that they would be given an opportunity to practice the interaction with emails before the actual tasks started. They had two chances to get the task correct or they would be excluded from the study. This task was not meant to be a screening task but to ensure a basic level of familiarity with the environment. The really basic level with a simple question was to ensure the measured effect was based on the video and not other factors. From 417 participants 121 failed the task and therefore could not proceed with the study.

Finally, all participants received a short overview of the next steps.

At this point, the next step depended on the assigned group (see also Fig. 5): The non-video groups proceeded directly to the interaction exercise. The video-groups watched the video, answered the questions, and then were redirected to the interaction exercise.

Anti-phishing video The video groups saw the five-minute video described in section 4.1. They could start the video themselves. They were then asked four questions about the video which we used as attention questions. These can be seen in the supplementary material. To proceed with the study, at least three of the four questions had to be answered correctly. Participants who answered fewer than three out of four attention questions correctly were excluded. In total 12 participants were excluded due to this.

Interaction exercise All groups went through an interaction exercise in which they were asked to count links in emails. To do so, they were told that links could be integrated as text, buttons and/or images/logos. The participants could hover over the link with their cursor and either a status-bar or a tooltip would appear, depending on the study group. Regardless of the interface, the new element would display the URL behind the link. Participants who failed this task for the first email were given a second email to try. Participants who failed both email tests were excluded. See the supplementary material for these emails. They were also told that links in the study were deliberately deactivated.

Main task: judging emails Each participant saw sixteen interactive screenshots of emails in the Gmail environment (see section 4.3 for more information about the emails). These were displayed randomly. These screenshots were interactive and worked exactly like the ones used in the Interaction Exercise step. Each screenshot was presented on a separate page with two question: (1) This email is a... phishing email or legitimate email, and (2) How certain are you (on a Likert scale from 1 to 7)?

Feedback Everyone was asked how they proceeded to judge the emails in an open text question.

Demographics Every participant who received the information about the browser update received a debriefing that this was not actually the case. The participants saw a series of demographic questions, recording their gender, age (in age ranges) and experience with Gmail.

The initial phase ended by thanking the participants and providing a code they could use to claim payment (see section 4.4).

4.2.2. Retention phase

After five months, those participants who successfully completed the initial study phase were contacted to request participation in the retention phase experiment. All four groups were recalled, including those who did not see the video. This was so that we could check whether the results were similar to before, allowing us to exclude external events that caused a shift in the detection ability of all participants. The rationale is that if something caused an increased awareness in the general population, this would be mirrored in a higher performance five months after the first evaluation across all participants.

The specific steps of the study (depicted in Fig. 5) played out in the same way as during the initial study phase with three differences: (i) the video groups were not shown the video again, but were referred directly to the interaction exercise. (ii) The exercise no longer applied exclusion criteria, but was only a reminder of how the interface worked. As the participants had to pass the exclusion criteria during the initial study phase to be able to complete the study, we did not find it necessary to use these again. We did not provide participants with feedback on their performance. This is addressed in section 6.7. The specific questions in the steps Interactive Exercise, Judging Emails and Feedback are provided in Appendix B (see Fig. B.6), Appendix C (see Fig. C.7) and Appendix D (see Fig. D.8).

4.3. Interactive email screenshots

There are various URL phishing techniques that aim to trick and confuse users while analyzing a link. Thus, before describing how the email screenshots themselves were selected, we introduce the URL phishing techniques we considered.

4.3.1. URL phishing techniques

Different papers (e.g., Mossano et al. (2022); Petelka et al. (2019); Reynolds et al. (2020)) consider different phishing techniques, but all consider the four types we list in this subsection – this also holds for Volkamer et al. (2018).

Obfuscate An arbitrary domain name or IP address is used to hide the destination. The URL lacks a connection to the (faux) sender of the email content. For example, in a phishing Amazon email, the URL behind a link is either "`www.host745.com`" or "`https://87.147.12.250`".

Mislead The name of the supposed sender company is used either in the subdomain area or in the path following the domain. For example, in a phishing Amazon email, the URL behind a link is either "`www.amazon.de.host745.com`" or "`www.host745.com/www.amazon.com`".

Mangle The name of the supposed sender company is used in the domain but with small, subtle changes. For example, in a phishing Amazon email, the URL behind a link is either "`www.amazno.com`" (two characters inverted) or "`www.arnazon.com`" (using r n instead of m).

Delusive mismatch URL The link text resembles a URL (e.g., "`www.amazon.com`" or "`amazon.com`"). The link text matches the domain-top-level-domain combination of the supposed sender company, but the URL behind the link directs the users to a different location. This technique can be combined with any of the previous three for a greater effectiveness. However, to avoid adding difficulties unrelated to the

Table 3
Overview of the phishing techniques matched to the companies. (4) Link text in email: https://brief.gmmx.net/AGB. (5) Link text in email: https://www.paypall.com/gutschein. (6) Link text in email: https://premium.gmx.de/speichervoll. (7) Link text in email: https://www.paypal.com/.

Company	Strategy Types	URL
Amazon	Obfuscate	https://telefon.host745.com/hinzufuegen
Lufthansa	Obfuscate	https://87.147.12.250/buchungs%C3%A4nderung]https://87.147.12.250/buchungsänderung
Google	Mislead	https://www.google.com.megahoust.ru/sicherheitscheck
LinkedIn	Mislead	https://login.linkyzt.com/www.linkedin.com/profil
DHL	Mangle	https://account.dlh.com/zustellung
Netflix	Mangle	https://www.netfllx.com/neuerlogin
GMX	Delusive mism. URL	https://premium.host547.ru/speichervoll[6]
PayPal	Delusive mism. URL	https://www.hokpurt.ru/AGB[7]
Amazon	Legitimate	https://packet.amazon.de/paketverfolgung
Lufthansa	Legitimate	https://www.lufthansa.com/buchungsanzeige
Google	Legitimate	https://www.google.com/neuesger%C3%A4t]https://www.google.com/neuesgerät
LinkedIn	Legitimate	https://video.linkedin.com/kurs
DHL	Legitimate	https://mailing.dhl.de/wunschort
Netflix	Legitimate	https://www.netflix.com/neuepreise
GMX	Legitimate	https://bestaetigung.gmx.de/AGB[4]
PayPal	Legitimate	https://www.paypal.com/gutschein[5]

specific phishing technique, we used the delusive mismatch URL in combination with the obfuscate technique. For example, in a phishing Amazon email, the link text reads "`www.amazon.com`" but the URL behind the link directs to "`www.host745.com`".

An overview of the specific URLs for each of the four types used in the study is shown in Table 3.

4.3.2. Study email screenshots

Similar to previous studies Berens et al. (2022); Canfield et al. (2015); Mossano et al. (2022); Reinheimer et al. (2020), we used an equal number of phishing and legitimate emails. All phishing techniques listed in the previous subsection were covered twice in the study, leading to eight phishing email screenshots and eight legitimate email screenshots, hence sixteen email screenshots overall. We chose to have two email screenshots per phishing technique to reduce the probability that participants' biases (e.g., personal opinions on the organization) triggered incorrect assumptions regarding the phishing technique employed in the email. As stated in section 4.2, these screenshots were interactive and would display the URL behind the link once the latter was hovered with the mouse cursor. The displayed interface depended on the participant's study group (status-bar or tooltip).

As basis for the email screenshots, we used real-world emails from well-known companies. We assumed that the phisher was cloning these and replacing the URLs, i.e., being able to spoof the "from" email address.

The study was conducted in German with German participants. Thus, we chose highly popular companies in Germany (see Table 3), so that users would not reject the emails based on their unfamiliarity with the sender. We also wanted to eliminate the influence of company reputation on the decision and, thus using both phishing and legitimate email screenshots for every company – with two different emails as harnesses.

4.4. Ethics, recruitment and payment

In this section we describe how we recruited our participants, how we identified them between the study phases and the payment they received.

4.4.1. Ethics

The study description was submitted for consideration and approved by the ethical board of our university. Note that the data protection is an integral part of the ethic submission. In addition to the ethical board, also the data protection officer of our university reviewed and approved both the informed consent and the overall study design.

4.4.2. Recruitment

We recruited 193 participants through the panel service "Clickworker", limiting the selection to those from Germany.

The initial phase started in December 2021 and the retention phase at the end of May 2022.

4.4.3. Participants tracking

Our study design required us to track participants through two different phases, five months apart. However, we also wanted to respect their anonymity. To solve this, we used the ID code assigned to each participant by the Clickworker panel service. We stored only the ID (and no IP address) alongside their performance, allowing us to connect performances from the same participant across databases. No sensitive information was collected, as only the panel service could associate each ID code to the associated person.

4.4.4. Payment

The initial study phase should have taken around 30 minutes to complete, based on our pre-tests. As we wanted to pay our participants at or above the German minimum wage, we calculated the payment according to the latest minimum wage at the time, i.e., € 9.82/h at December 2021, rounded up to € 10/h. However, because the video groups had to spend more time in the study, we also calculated the payment on the longest group, as participants were assigned randomly to the groups. All previous points considered, the final payment for the initial study phase was € 10/h * 30 minutes = € 5.

Albeit the retention phase would have taken less time, considering that no video was shown, we still decided to pay the participants € 5 to reward their willingness to return for the retention phase. At the time of the retention study, this still corresponded to half an hour working at minimum wage.

5. Results

In this section, we present the results with respect to the different research questions and hypotheses.

5.1. Data cleaning

193 participants finished the first survey and were invited to the retention phase. 82 of these 193 participants completed the online survey for the retention phase. These numbers are similar to other papers like Mayer et al. (2014) and Volkamer et al. (2018), that have around 50% dropout rates in retention studies. In our case, the dropout may have been higher as we used a panel service, while related work recruited via social media, leaflets, and word of mouth. After five months, several of the 193 Clickworker accounts (about half of those who did not

Table 4
Usage distribution of Gmail for all four groups.

	Using Gmail	Used Gmail	Never used Gmail
StatusBarVideo	18	1	5
StatusBar	13	4	3
TooltipVideo	11	1	3
Tooltip	14	1	5

participate in the retention phase) were not active anymore, thus the invitation was not received as they only receive invitations through the platform itself.

For the remaining 82, we performed the following data cleaning step: We calculated outliers and excluded those that violated the maximum of 1.5x interquartile range (IQR). More on the testing for assumptions in the next section. Therefore, we excluded three participants. Thus, we analyzed data from 79 participants to consider our research questions: StatusBarVideo: 24, StatusBar: 20, TooltipVideo: 15, Tooltip: 20.

5.2. Demographics

This subsection, provides an overview of the main demographics of our participants. In cases of differences between the groups, we tested whether this had an effect on our main research questions. An overview of the age distribution is provided in Table 6. Middle age groups are slightly over represented, especially in the StatusBarVideo group. There are also more male than female participants in all groups (see Table 7). We checked the differences between male and female participants over all four groups (see Table 5). Using a two-samples Wilcoxon test, we found a significant difference between female and male participants for the initial phase ($W = 867.5,\ p = 0.041$). Looking at the descriptive data, there seems to be no gender difference for the video groups, at least for StatusBarVideo with similar participants, while there seems to be difference between groups without a video, e.g., Tooltip and StatusBar.

The groups were evenly populated by participants with and without knowledge of Gmail (see Table 4). Given the small numbers and knowledge of Gmail, we decided not to check if this influenced their performance.

5.3. Descriptive statistics and mixed ANOVA results

The results for phishing detection rate per group and phase are shown in Table 8.

To answer RQ1-RQ4 and RQ10, we conducted a mixed ANOVA for the between-subject factor "group" and the within-subject factor "phase".[7] First, we checked the parametric assumptions and found that only homogeneity of covariances was breached. Because Field Field (2013) points out that ANOVA tests are relatively robust against this breach, we continued the analysis, but omitted the interpretation of the interaction term. The results show there are statistically significant effects on phishing detection ($F(3, 75) = 4.22,\ p < 0.0001$) for both the between-subjects "group" factor ($p < 0.0001,\ \eta_g^2 = 0.27$) and the within-subjects "phase" factor ($p < 0.0001,\ \eta_g^2 = 0.03$). The post-hoc tests are reported in the sub-sections of the respective RQ.

5.4. Pre-requisite research questions results

We present here the analysis and the results to our pre-requisite research questions.

5.4.1. ***RQ1/H*$_1$**: *TooltipVideo effectiveness*

We used a post-hoc t-test to test for differences in phishing detection rates. Considering the Bonferroni adjusted p-value, we found significant differences for the tooltip groups ($p < 0.0001$). The group TooltipVideo ($mean_{TooltipVideo} = 89.3\%$) outperformed the group Tooltip ($mean_{Tooltip} = 63.5\%$). Accordingly, we accepted $\mathbf{H}_1$ that watching the video leads to a significantly higher ability to detect phishing emails for the tooltip interface.

5.4.2. ***RQ2/H*$_2$**: *StatusBarVideo effectiveness*

We used a post-hoc t-test to test for differences in phishing detection rates. Considering the Bonferroni adjusted p-value, we found significant differences ($p < 0.0001$). The group StatusBarVideo ($mean_{StatusBarVideo} = 81.2\%$) outperformed the group StatusBar ($mean_{StatusBar} = 49.5\%$). Accordingly, we accepted $\mathbf{H}_2$ that watching the video leads to a significantly higher ability to detect phishing emails for the status-bar interface.

Both pre-requisites were met, i.e., the video had a significant effect on the ability of participants to detect phishing emails. Hence, we could proceed in our investigations of the main research questions.

5.5. Main research questions results

We present here the analysis and the results of our main research questions.

5.5.1. ***RQ3/H*$_3$**: *TooltipVideo effectiveness after 5 months (retention)*

We use a post-hoc t-test to test for differences in detection rate of phishing emails. We found significant differences ($p = 0.0006$). The group after five months[8] that watched the video ($mean_{TooltipVideo} = 82\%$) outperformed the group without a video ($mean_{Tooltip} = 63.5\%$). Accordingly, we accepted $\mathbf{H}_3$ that watching the video leads to a significantly higher ability to detect phishing emails for the tooltip interface five months later.

5.5.2. ***RQ4/H*$_4$**: *StatusBarVideo effectiveness after 5 months (retention)*

We used a post-hoc t-test to test for differences in detection rate of phishing emails. We found no significant differences ($p = 0.1957$). Accordingly, we rejected $\mathbf{H}_4$ that watching the video leads to a significantly higher ability to detect phishing emails for the status-bar interface five months later.

5.5.3. ***RQ5**: How decided*

Two coders independently applied deductive coding to analyze the answers on how they decided whether it was a phishing email or not. First, two authors developed a codebook from the answers from all four groups in both phases. The codebook contained the following codes (which describe the different aspects of the email the participant draws attention to – in particular, if the link was the only reliable factor or not):

- "focus on link related aspects": Participants stated to only evaluate link related aspects and did not mention anything else. Note, it was not necessary to describe in detail what exactly was checked.
- "mentions links related aspects plus other aspects": Participants stated they checked the link but in addition mentioned other cues they checked for (e.g., grammar, design, sender).
- "nth link related": Participants stated various cues they checked for, but nothing link related.
- "others": There were two reasons for coding a statement as others: Participants talked about their difficulties during the email evaluation *or* the message of the answer was not clear to the coder. It was possible to assign this code on top of one of the first three codes.

[7] Please note, for the remaining research questions, we use different tests and report them in the corresponding subsections.

[8] The specific time-interval data are in Appendix A.

Table 5
Percentage of correct phishing answers per gender overall and for all four groups.

	Overall		TooltipVideo	StatusBarVideo	Tooltip	StatusBar
	Mean	SD	Mean	Mean	Mean	Mean
Female	64	26.3	87.5	83.8	51.6	43.8
Male	76.6	23.9	94.6	82.1	70.8	58.0

Table 6
Age-groups distribution for all four groups.

Group	18-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	>65
StatusBarVideo	1	1	7	3	1	2	2	1	3	3	0
StatusBar	0	3	3	1	4	2	1	2	3	1	0
TooltipVideo	1	0	3	2	2	1	1	0	2	1	2
Tooltip	0	1	1	2	3	6	4	1	0	2	0

Table 7
Distribution of the gender for all four groups.

Group	Male		Female	
StatusBarVideo	14	58.33%	10	41.67%
StatusBar	14	70.00%	6	30.00%
TooltipVideo	14	93.33%	1	0.63%
Tooltip	12	60.00%	8	40.00%

Table 8
Mean percentage and standard deviation of correctly detected phishing and legitimate examples divided both by study group and by time of measurement.

	Group	Initial		Retention	
		Mean	SD	Mean	SD
Phishing	StatusBarVideo	81.20%	18.00	60.40%	32.10
	StatusBar	49.50%	20.40	46.00%	24.60
	TooltipVideo	89.30%	9.61	82.00%	12.10
	Tooltip	63.50%	22.10	62.50%	20.70
Legitimate	StatusBarVideo	82.50%	20.10	79.20%	17.40
	StatusBar	75.20%	22.50	76.70%	22.90
	TooltipVideo	80.60%	20.10	81.80%	13.80
	Tooltip	83.00%	15.30	84.50%	16.40

- "sender": Participants mentioned that they checked the sender information (while it could be just sender or sender address or sender email address). It was possible to assign this code on top of the second and the third code.

The codebook was used for the deductive coding by two authors. Depending on the answer, more than one code was assigned. Please note, we only coded the answers from those participants who also took part in the retention phase. The authors agreement on the coding was then confirmed by calculating both the percentage agreement (95.43%) and Cohen's k (0.89). According to Cohen (1968); McHugh (2012), values of $k > 0.81$ and $PercentAgreement > 95$ indicate very good to perfect agreement. Hence, the values achieved are in an acceptable range and can be used for further interpretation. For the initial phase we got on average 23.38 words with a median of 18 and for the retention phase we got 22.99 words with a median of 18. The distribution is skewed towards short answers with $skewness_{intial} = 1.36$ and $skewness_{retention} = 2.01$. Table 9 shows which group mentioned which code and how often. In particular for the StatusBarVideo group, less participants mentioned that they only focused on checking the link (the number decreased from 15 to 6) and more people stated nothing link related anymore (the number increased from 0 to 7). The numbers also indicated that the status-bar interface made it less likely that participants checked the link.

5.5.4. *RQ6: Refreshments*

For this section, we looked at the phishing detection rate of the individual phishing techniques during the retention phase (see Table 10 for the different techniques and groups). Our focus for this research question was the TooltipVideo group. The goal was to find out if it is needed to refresh all the content from the video or if it is sufficient to focus on some phishing technique, as it's the only one to perform significantly better after five months. The three phishing techniques Obfuscate, Mislead and Delusive mism. URL achieved 90% correct answers or more after five months. The phishing technique Mangle only achieved around 57% after five months.

5.5.5. *Further research questions results*

As mentioned in section 3.3, we were also interested in answering further research questions connected to our main ones. Here we present the analysis and the results of them. For RQ7 and RQ8 we conducted a repeated measure ANOVA and followed the same approach as for RQ1-RQ4 and RQ10.

5.5.6. ***RQ7**: Tooltip vs. StatusBar - influence of video*

When looking at the numbers for phishing detection in the initial phase (see Table 8), we saw that the difference for detection rate for the StatusBarVideo group is on average 31.7% higher (from 49.5% to 81.2%) compared to the TooltipVideo group with 25.8% (from 63.5% to 89.3%). So related to the research question, we found a larger effect from watching the video for the StatusBarVideo group than for the TooltipVideo group. As all data for all groups are from different individuals there is no possibility to link data from the groups without video to their counterpart with video and therefor testing the differences for significance.

5.5.7. ***RQ8**: Tooltip vs. StatusBar - influence of video (retention)*

When looking at the numbers for phishing detection from after watching the video to the retention phase (see Table 11), we found that detection rate for the TooltipVideo group decreased on average less with -7.3% compared to the StatusBarVideo group -20.8%. So related to the research question, we found a larger effect after five months for the TooltipVideo group.

5.5.8. ***RQ9**: Certainty - influence of video*

To answer this RQ, we looked at the descriptive values for certainty (reported in Table 12) and compared the different groups. Participants were asked, on each email, to state how certain they were about the decision they made. The scale ranges from 1 (very uncertain) to 7 (very certain). The mean certainty value was above the neutral range (i.e., value of 4) for both the correct decision and the incorrect decisions - for all groups. The effect of the video on participants certainty (independently from whether the answer was correct or not) was larger for the StatusBar than for the Tooltip group. The effect of the video on participants certainty for emails they correctly identified as legitimate emails was slighter larger for the Tooltip group. The effect of the video on participants certainty for emails they correctly identified as phishing emails was slightly larger for the StatusBar group. Comparing the

Table 9
Frequencies of applied codes for each group during initial and retention phase. In general, the higher the number in the first two codes the better (in particular for the first one). For the code 'nth link related, the smaller the number the better). The numbers in brackets are those for participants who did not come back in the retention phase. Sender is coded jointly with other ones, hence it is calculated on its own.

	Group	Focus on link related aspects	Mentions links related aspects plus other aspects	nth link related	Others	Sender
Initial	StatusBarVideo	62.50% (47.62%)	37.50% (47.62%)	0% (0%)	0% (4.76%)	29.17% (33.33%)
	StatusBar	6.67% (17.14%)	40.00% (65.71%)	16.67% (14.29%)	3.33% (2.86%)	55.00% (42.86%)
	TooltipVideo	43.75% (46.67%)	56.25% (43.33%)	0% (0%)	0% (10.00%)	31.25% (23.33%)
	Tooltip	31.82% (10.71%)	63.64% (57.14%)	4.55% (28.57%)	0% (3.57%)	36.36% (57.14%)
Retention	StatusBarVideo	54.17%	29.17%	16.67%	0%	50.00%
	StatusBar	9.52%	52.38%	33.33%	4.76%	61.91%
	TooltipVideo	26.09%	34.78%	4.35%	0%	34.78%
	Tooltip	47.62%	52.38%	0%	0%	28.57%

Table 10
Mean and standard deviation (SD) of correct answers (in percentage) of the four phishing techniques per group and per phase. Bold highlight for the group that is considered for RQ10 and star for the phishing technique that performed the worst.

	Group	Initial		Retention	
		Mean	SD	Mean	SD
Obfuscate	**TooltipVideo**	**96.67%**	**12.91**	**96.67%**	**12.91**
	Tooltip	80.00%	34.03	70.00%	29.91
	StatusBarVideo	89.58%	29.41	62.50%	39.70
	StatusBar	65.00%	28.56	72.50%	30.24
Mislead	**TooltipVideo**	**100.00%**	**0.00**	**90.00%**	**20.70**
	Tooltip	65.00%	32.85	67.50%	40.64
	StatusBarVideo	87.50%	22.12	62.50%	39.70
	StatusBar	60.00%	38.39	47.50%	41.28
Mangle	**TooltipVideo**	**80.00%**	**31.62**	***56.67%****	**41.69**
	Tooltip	25.00%	38.04	30.00%	41.04
	StatusBarVideo	58.33%	43.41	39.58%	38.95
	StatusBar	32.50%	37.26	17.50%	29.36
Delusive mism. URL	**TooltipVideo**	**100.00%**	**0.00**	**96.67%**	**12.91**
	Tooltip	82.50%	33.54	92.50%	24.47
	StatusBarVideo	95.83%	20.41	81.25%	38.48
	StatusBar	57.50%	43.76	55.00%	48.40

Table 11
Difference for legitimate and phishing detection rate for all four groups. Additionally the difference between the initial phase and the retention phase.

	Group	Initial Study Mean	Retention Mean	Difference
Phishing	StatusBarVideo	81.20%	60.40%	-20.8%
	StatusBar	49.50%	46.00%	-3.5%
	TooltipVideo	89.30%	82.00%	-7.3%
	Tooltip	63.50%	62.50%	-1.0%
Legitimate	StatusBarVideo	82.50%	79.20%	-3.3%
	StatusBar	75.20%	76.70%	1.5%
	TooltipVideo	80.60%	81.80%	1.2%
	Tooltip	83.00%	84.50%	1.5%

certainty for wrong answers with that for correct ones, we saw that the certainty for correct answers was only slightly higher.

5.5.9. *RQ10/H_{10}: Tooltip vs. StatusBar - general performance*

The descriptive data to answer this research question is provided in Table 8. In section 5.3, we report that there are significant differences between groups and phases with a large effect of $\eta_g^2 = 0.27$ for "group". We used a post-hoc test to test for differences in phishing detection rates. Considering the Bonferroni adjusted p-value, we found a significant difference between the two groups of participants that did not watch the video ($p = 0.0395$). Accordingly, we accepted the $\mathbf{H}_{10}$ that the tooltip interface has a greater positive effect on phishing detection than the status-bar.

Note, there was no significant difference between the groups of participants who saw the video ($p = 0.19$) - while the StatusBarVideo group had lower rates compared to the TooltipVideo group (see Table 8). The

Table 12
Correct respectively wrong answers for legitimate and phishing emails for the four different groups. The scale ranges from 1 (very uncertain) to 7 (very certain) with the own decision on the specific example to be judged.

Group			Initial		Retention	
			Mean	SD	Mean	SD
StatusBarVideo	Phishing	Correct	6.45	0.67	3.31	0.82
		Wrong	5.61	0.91	3.43	1.06
	Legitimate	Correct	6.10	0.74	3.73	0.86
		Wrong	5.04	1.88	2.74	1.60
StatusBar	Phishing	Correct	5.29	1.02	3.21	1.18
		Wrong	5.53	1.06	3.50	0.84
	Legitimate	Correct	5.60	0.68	3.09	0.89
		Wrong	4.92	0.89	2.68	0.92
TooltipVideo	Phishing	Correct	6.45	0.55	3.71	0.68
		Wrong	5.14	2.04	3.25	0.94
	Legitimate	Correct	5.71	0.91	3.09	0.88
		Wrong	5.48	1.24	3.35	1.49
Tooltip	Phishing	Correct	5.70	1.08	3.22	0.91
		Wrong	5.03	0.89	2.73	1.21
	Legitimate	Correct	5.56	0.66	3.08	0.96
		Wrong	4.76	1.19	3.57	1.55

Table 13
Distribution of participants being informed about phishing prior to the study per group. Combined with their certainty for phishing, legitimate and all emails. Also with the detection rate for phishing emails.

Group	Previous Information	Informed	Certainty Phish	Certainty Legitimate	Certainty Overall	Detection Rate Phish
StatusBarVideo	Yes	66.67%	6.59	6.18	6.39	83.80%
	No	33.33%	5.73	5.48	5.61	76.20%
StatusBar	Yes	70.00%	5.45	5.41	5.43	52.90%
	No	30.00%	5.29	5.48	5.38	41.70%
TooltipVideo	Yes	80.00%	6.46	5.82	6.14	90.00%
	No	20.00%	6.04	5.04	5.54	86.70%
Tooltip	Yes	70.00%	5.62	5.51	5.57	65.70%
	No	30.00%	5.06	5.33	5.20	58.30%

factor "group" from the Mixed ANOVA also represented a large effect, with $\eta_g^2 = 0.270$ (see Section 5.3). Accordingly, these comparisons of the groups could also become significant with an larger sample size.

5.6. Further results

Here we present further results that were not formulated as research questions.

5.6.1. Influence of previous anti-phishing awareness

We asked participants if they had previously informed themselves about phishing. At least 66.6% of the participants per group had done so (in the TooltipVideo group 80%, Tooltip group 70%, StatusBarVideo group 66.6%, and in the StatusBar group 70%). Regarding the anti-phishing awareness measures types, participants answered the following[9]: the Internet, specifically phishing warnings about recent phishing waves rather than general information (30), work (12), news paper / news article (10), acquaintances, such as friends, colleagues, home (6), school (3), university (3), and bank (1).

We were interested in gaining more insights into the difference between the group of participants who stated that they received phishing information before the study and those who stated they did not receive any – while focusing on the two study groups StatusBar and Tooltip.

There seemed to be no difference in terms of certainty between those who already had information prior to the study ($mean_{Tooltip} = 5.62$, $mean_{StatusBar} = 5.45$) compared to those that did not have information ($mean_{Tooltip} = 5.06$, $mean_{StatusBar} = 5.29$). Furthermore, there seemed to be a difference in the detection rate of phishing emails between those groups. For those who already had information prior to the study the detection rates were 65.7% for the Tooltip group and 52.9% for the StatusBar group compared to those groups who stated that they did not receive some (with 58.3% for the Tooltip group and 41.7% for the StatusBar group).

5.6.2. Influence of the company

Although all emails were authentic and only a single aspect changed (the URL), we wanted to check whether some were judged as phish more often than others. For the analysis, we counted both the emails correctly judged as phishing and the legitimate emails wrongly judged as phishing (see Table 14). For example, Amazon emails ($mean = 63\%$) were judged as phish more often than those from Netflix ($mean = 25\%$). This means that, regardless of whether the example was a phish or legitimate, 63% of Amazon emails were judged as phish against only 25% of Netflix emails.

5.7. Legitimate emails and uncontrolled effects

Additionally to the phishing detection rate, we also checked the rate of correctly identified legitimate emails for every research question. This was done because if a significant higher number of legitimate

[9] Please note, some participants mentioned more than one type of measure.

Table 14
Percentage of emails of a company judged as phishing (both legitimate and phishing).

	Initial	Retention
Amazon	62.66%	59.49%
GMX	56.33%	55.70%
Apple	55.70%	50.00%
PayPal	55.06%	55.06%
LinkedIn	43.67%	35.44%
Lufthansa	43.04%	38.61%
Google	41.14%	36.08%
Microsoft	36.08%	32.91%
DHL	32.91%	21.52%
Netflix	24.68%	22.15%

emails were judged as phish after watching the video than before, it could have meant that participants got overcautious from the video and did not actually detect more phishing emails as such. However, we found no such trend in our data. The test results are in Appendix A.2.

Furthermore, we checked for uncontrolled effects that might have introduced differences in the general population during the time interval. We did this to make sure that an uncontrolled effect in the general population, like a highly publicized phishing attack, did not influence everyone's awareness and, consequently, the retention data. To check this, we compared the StatusBar phishing detection from the initial phase to that of the retention phase, as any performance difference would then not have been caused by the video. We found no such effect. Hence, we can exclude that external factors created a phantom effect in the general population. For the test results, see Appendix A.3.

6. Discussion

We first discuss our main findings and then acknowledge the limitations of our study.

6.1. Video effectiveness at retention

The results of RQ3 indicate that the TooltipVideo outperformed the Tooltip after 5 months. The performance of both video groups appears to be similar to that of the related work discussed in section 2: After watching the video, the participants reached a phishing detection rate between 81.2% (status-bar) and 89.3% (tooltip). However, RQ3 and RQ4 results indicate that, at the retention phase, the TooltipVideo group's performance (82.00%) seemed superior to that of the StatusBarVideo group (60.4%). The TooltipVideo group appears to be in line with the results of related work (with the exclusion of Reinheimer et al. (2020), 71%[10]).

This result suggests that shorter measures such as our video could be as effective for tooltips both right after the measure and after five months (see Table 8) as more time-intensive measures such as the game evaluated in Canova et al. (2015b) and the e-learning used in Berens et al. (2022) (reported in Table 2; note, the data aggregate status-bar and tooltip results).

However, a five months retention is still limited, as this means one would need to refresh knowledge after half a year already (as recommended by Reinheimer et al. (2020) based on their results that were not significant anymore after six months). Thus, the question is if there is a way to extend the measure so that the effect is still significant after more than 6 months, maybe aiming for a year. According to Beyer et al. (2015) and Sasse et al. (2022) internalizing secure practices are required to truly reach a shift towards secure behavior and thereby achieving a long lasting effect. One way to implement this approach while keeping the anti-phishing awareness measure short could be to combine the video with a small challenge or quiz in which judging emails are practiced. As future work, the long-term effect of such a combination should be studied.

6.2. Performance status-bar vs. tooltip

Our results for RQ10 regarding the *non-video groups* suggest that a status-bar is less effective than showing in addition a tooltip, with regard to phishing detection. This finding is inline with the results from Petelka et al. (2019) (the closer the URL is to the link, the more effective it is) and Volkamer et al. (2016) (showing that a tooltip is more effective in supporting people to detect phishing emails than a status-bar). Admittedly, the aforementioned work focus on different aspects than comparing status-bar and tooltip. Nonetheless, as one of their main results is that just-in-place elements have greater effects on users' phishing detection, and because a tooltip is a just-in-place element, we believe it is an acceptable inference.

Althought the StatusBar group performance is lower than the Tooltip group one, looking at the results for RQ1 and RQ2, we can see that the StatusBarVideo group reached a similar level of phishing detection to the TooltipVideo group. Hence, as shown in RQ7 results, the video effect is greater for the status-bar than for the tooltip (+31.7% vs. +25% over the baseline). This might be caused by fewer participants being aware of the status-bar existence and function in the first place, leading to a higher number of them being informed by the video.

The results from RQ4 suggest that the StatusBarVideo group is no longer significantly better than the StatusBar after 5 months. This result appears to confirm that effective security protection can only be reached if (1) users are aware of the threats and how to protect against them, and (2) if the security mechanisms for this protection are usable and cannot be easily missed - which is not the case for the status bar. Also the results from RQ5 suggest that the status bar interface makes it less likely that participants based their decision on the link after five months.

Furthermore, indications of the apparent lower performance of the status-bar come from other results too. If we consider the results from RQ8, the tooltip location seems to have some influence on how long the knowledge acquired remains useful. Even though the initial video effect is greater on the StatusBarVideo group, during the retention phase the TooltipVideo group suffers a lower loss of knowledge than the StatusBarVideo one (-7.3% vs. -20.8%). A possible explanation of this might be that the just-in-place characteristic of the tooltip allows it to be much more noticeable, acting as a sort of ever present reminder to look at the URL and reducing the degree of knowledge decay. The status-bar, being out of the participants' center of attention, it is much easier to forget and not consider.

From all of the above, then, it seems that showing a tooltip with the status-bar leads to better phishing detection.

6.3. Problems with current awareness measures

The results of RQ5 (in Table 13 in section 5.5.3) show that 70% of the participants were already knowledgeable about phishing, receiving information from various sources. This suggests that obtaining information on phishing is not such an uncommon fact.

According to the results of RQ5, the answers participants gave in the video-groups at the retention phase suggest a return to strategies based on unreliable cues, i.e., not based on link analysis. A possible explanation might be the lack of quality of existing phishing awareness measures (as identified by Mossano et al. (2020)). If in the past they were told several times that checking for spelling, emotions and (maybe) the sender is important (but nothing on the link), they may be more likely to remember these recommendations rather than the link one which they were (potentially) only informed about once – by our video.

[10] One reason for the lower mean rate in Reinheimer et al. (2020) could be that the awareness measure also covered two further security topics.

RQ5 results also show that the certainty of the non-video groups was above neutral and almost on par with the video groups (reported in Table 12). However, when we compare information received, the declared confidence, and the actual phishing detection rate, we notice a discrepancy: Most participants seem to overestimate their phishing detection ability. This result appears to confirm the data from Wang et al. (2016) on overconfidence.

A possible explanation for the observed overconfidence might lay in the results of Mossano et al. (2020): The fact that many participants receive information might not mean that the information is correct, nor that it is fitting for everyone. Different sources might give users different (if not conflicting) recommendations, leading to a false sense of security. In turn, this overconfidence might lead to the misplaced certainty we found.

The overconfidence might lead participants to the belief that they have no need to receive further awareness, creating a state of reluctance to search for further recommendations. In turn, this reluctance might lead to not change the sub-optimal phishing detection strategies, putting users at increased risk of falling for phishing attacks. One possible solution to this issue is exploring users' reactions when presented with the distance between their confidence and their actual performance, as it might lead to a shift in their mental model. Another solution is to pursue the proposal by Mossano et al. (2020) to achieve a greater level of standardization among different awareness measures. As future work, we plan to follow both lines of research.

6.4. Recommendations for the refreshment

According to the results of RQ6, the mangle phishing technique appears to be the most difficult technique to notice (see Table 10). Therefore, our recommendation is to not necessary repeat all the phishing techniques (to save time) but focus on the mangle phishing technique. The mangle might be the most difficult technique, because it needs the full attention of participants. Also depending on their environment, small differences can be really hard to spot. Moreover, humans recognize words by their shape and letter order, so, even if such tricks are known, it is difficult to overcome this process. Focus on the mangle phishing trick could be achieved by adding information to the video about the difficulty for the brain to notice the spelling error in the domain name if they do not just focus on the URL, but the entire authentic looking email. Additionally making people more aware of this special trick and how easy it is to do not see the spelling errors e.g. given an example text were the characters are misplaced, but the human brain is still able to read the word as a whole. Additionally tricks such as increasing the font size and try reading character by character. Such recommendations are similar to functionalities that tools provide trying to address these phishing tricks by putting spaces between characters.

6.5. Future works on our video

It might be worth evaluating as future work approaches different than simply stating that it is important to check the link: A potential approach might be to show users how easily legitimate emails can be cloned, i.e., modified with a phishing URL behind the embedded links (techniques described in Pienta et al. (2018)). The idea would be to demonstrate to users, with practical examples, why their current strategies are not effective against advanced phishing technique.

Our video focuses on the link because the URL behind it is the only reliable factor to decided whether it is risky to click on a link or not. However, many of today's phishing emails can be detected by checking the sender (and the spelling/emotions). Thus, another approach could be to integrate these checks in the refreshment measure. First, check the sender information and whether the spelling or the emotions are suspicious. If this is not the case check the link before clicking on it. The goal would be to integrate the link checking in people's security routines (as recommended by Beyer et al. (2015) and Sasse et al. (2022)).

6.6. Comparison with related work

Another interesting outcome of our study is the apparent confirmation of the results from Albakry et al. (2020). They found that specific companies elicit different reactions in users when employed as phishing emails. The results in section 5.6 appear to confirm their findings. As it can be seen in Table 14, e.g., Amazon emails are twice as likely to be marked as phishing than DHL emails or Netflix ones. Please note, all emails used in our study were emails sent by the corresponding company to one of the authors. We only changed URLs behind links. The differences we observed for the companies might be caused by the prevalence of certain phishing hooks, i.e., Amazon emails might be perceived as more likely to be used in phishing attacks. It might also be that many Germans have an Amazon account and react differently because of that. As future work, this effect should be further studied as the second reason is less critical than the first one. The first one would mean that if they have not heard of, e.g., DHL being a target of phishing, they are less likely to check carefully.

Differently from the related work, we found no significant difference in participants' legitimate detection, with or without the video. This was also the case at retention, as the participants' performance barely changed between the two phases. The lack of significance was surprising, but not inexplicable: previous research might have revealed differences in the participants' legitimate performance because the non-interventions groups appeared to be worse than our participants.

A further result of our study – and in particular RQ5 – seems to be that we found that the most difficult phishing technique to recognize is the mangle one, as shown in Table 10. This is interesting, because in Reynolds et al. (2020), this was the easiest technique to notice. The difference might be due to the way in which the URLs where shown: Reynolds et al. (2020) show the URLs in isolation, i.e., not in a context. In our study, instead, the URLs are set in a use context, i.e., as links in emails. This might show that when studying the impact of URL phishing techniques, it is very important to place them in a use context, where there are other factors that could distract the participants, as it happens in every day life (see halo-effect Kirlappos and Sasse (2012); Nisbett and Wilson (1977)).

6.7. Limitations

We employ a tooltip interface similar to the status quo in Microsoft Outlook app v.16 on Windows. However, the interactive screenshots in our study are set in the Google Gmail environment in Chrome, which does not use a tooltip to display the URL, but rather a status-bar. This might have caused confusion to participants familiar with the Gmail interface, i.e., the majority of our participants (see section 5). We addressed this by informing participants in the tooltipgroup that they will see the interfaces with the newest updates. We decided not to explain what the update is about (i.e., we did not mention the tooltip) to not prime them. Note, web browsers and web pages change their interface from time to time without explaining what exactly is different. Thus, our framing can happen in the real world.

Besides issues with the content, the videos from Hamdani and Mustafa (2021) are proprietary videos, therefore we could not use them. In this context we acknowledge that the research is bound by some restrictions such as property of the right to use measures. Therefore results could differ when such right would have been acquired and at least one of the videos used for comparison. Still we think the issues with the content such as the lack of highlighting the statusbar and missing of the information about path and subdomain attacks, make the chosen video from Volkamer et al. more appropriate for the research.

Previous research on the same video from 2018 and the comparison with those results could be influenced by an increase in societal awareness. In our research we did not want to research such a societal awareness increase. Still we find it important to compare our results for some indication on how to interpret the results with the past. In

the research from 2018 around 60% of the phishing answers were correct. Comparing this to the current study, the status bar group achieves around 50% correct answers and the tooltip around 60%. In the previous study both were studied together. We acknowledge that there are differences in the examples used in the past and in the current study. Still, we have no indicator as to our examples being more difficult to judge, which would negate a general societal awareness increase. Therefore we see no indication for such an increase or influence on our results.

The interactive exercise might bias participants towards links, exposing that they play an important role in the study. However, we argue that the influence is negligible, as no information on URL phishing techniques are provided, only how links can be implemented in an email. Participants already know how to interact with links from their real-life experiences. Nonetheless, the interactive exercise presence is required to guarantee that participants understands that the email screenshots are interactive, i.e., that link can be hovered and a URL is shown in the interface of the specific group of assignment (either a tooltip of status bar).

Some participants may not have been aware that phishing detection has something to do with checking links. Yet, they were somehow informed about it through the interactive example to count links. Albeit this is not an issue for the video groups, as they would have been informed by the video anyway, it might have given knowledge to the participants in the non-video groups that they would have otherwise not possessed. We acknowledge that this might have influenced the results for the non-video group, but it was a necessary step to be sure participants were aware that we used interactive email screenshots. In any case, not providing such information to participants would further lower the performance of the none-video group. Hence, we believe that the study design actually helped the performance of the non-video groups, instead of hindering it.

Participants in the retention study saw the same screenshots they judged during the initial phase (similar to Berens et al. (2022), Reinheimer et al. (2020), and Volkamer et al. (2018)). We choose this instead of creating a new set of screenshots with the same qualities, i.e., neither more complex or with different contextual biases (e.g., the email layout), as the latter would've been very difficult. Yet, seeing the same screenshots again might have influenced their performance in the case they remembered their previous decisions. We did not ask them a control question for this variable, as both the long time between phases and the random order of the screenshots should have reduce the chance of this happening. Moreover, we also presented both a legitimate and a phishing version of each screenshot. Hence, even if they remembered one company's email as either legitimate or phishing, they would need to correctly remember the specific screenshot between the two possible ones.

Following the related works in section 2, we expected some steep decrease in the participants that returned for the retention. However, we incurred in an unexpected issue: as mentioned in section 4.4.2, we employed members of the panel service Clickworker as participants. When we contacted them again after five months, we found that almost 50% were no longer present on the platform. Therefore, our drop-out rate was much higher than expected. For future retention studies, especially long time online ones, we recommend researchers to overestimate their participants to control for the possibility of many of them not being part of the chosen panel service anymore.

We had no time constrain while judging the email screenshots. This is not realistic, as in a real-life setting, security would be a secondary goal of users, not their main one (as in our study design). However, given that our goal was to determine the knowledge retention of users watching our video and not a realistic representation of users interactions with emails, we do not believe the absence of time pressure impacted our research goal. However, it is important to evaluate the effect of time pressure too, hence we plan to investigate it in future studies.

As mentioned in section 4.2.2, we did not provide feedback to participants regarding their performance. We acknowledge that this is a shortcoming.

7. Conclusion

We improved our earlier research video on phishing and showed that it continues to be an effective measure to improve phishing detection – both in the short- and long-term. In addition, we gained further insights into people's phishing detection strategies and how they change after 5 months. Our results also enabled us to derive recommendations to develop reminder measures to be distributed after 6 months (note, 6 months as recommended by Reinheimer et al. (2020)). Tooltip interfaces seem to help knowledge retention better than only status-bar interfaces after 5 months. A single status-bar showed worse results in the beginning, greater improvements, but also a greater decline over the span of 5 months. Therefore these should be critically reevaluated in today's software. In addition, our findings suggest that there is already a subjectively high level of confidence concerning phishing detection - independently from having watched the video or having informed themselves about the topic in the past. This merits future investigations because, if our findings are confirmed, we need to research how to make people aware of the mismatch between their perceived and actual ability to detect phishing emails. Furthermore, multiple developed measures are freely available at present. However, after comparing them with our video, we found that they do not cover everything we cover. For the future, it would be interesting to cooperate with industry partners to evaluate and enhance their solution. This would also help to move towards greater standardization of the information given to users in improving phish detection. Finally, our clear recommendation is to avoid reading emails in contexts in which the information is only provided in the status bar as this makes it more likely that a phishing email will be missed.

CRediT authorship contribution statement

Benjamin M. Berens: Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft. **Mattia Mossano:** Conceptualization, Visualization, Writing – review & editing, Supervision. **Melanie Volkamer:** Supervision.

Declaration of competing interest

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

Data availability

The authors do not have permission to share data.

Funding sources

This work was supported by funding from the project "Engineering Secure Systems" of the Helmholtz Association (HGF) [topic 46.23.01 Methods for Engineering Secure Systems] and by KASTEL Security Research Lab.

Appendix A. Results

A.1. Retention period

Due to the time spans, the participants had several days to start the study for both phases. Therefore, the actual time passed between

Table A.15
Retention period for the four different groups with mean and standard deviation for days, weeks and months.

Group	Days		Weeks		Months	
	Mean	SD	Mean	SD	Mean	SD
StatusBarVideo	162.00	10.00	23.10	1.44	5.37	0.33
StatusBar	166.00	6.27	23.80	0.90	5.53	0.21
TooltipVideo	163.00	11.80	23.20	1.68	5.40	0.39
Tooltip	162.00	10.30	23.10	1.47	5.38	0.34

participating in the first phase and the second phase can vary. For an overview see Table A.15. Since all groups were only activated after five months, the average time is, as expected, slightly above the average number of weeks in 5 months. There are only 0.16 months difference in the actual retention interval between the shortest and the longest group, i.e., about five days (0.16 months * 30.437 days/month). Thus, we conclude that the groups do not differ much in the time between the actual initial phase and the actual retention phase.

A.2. Results for legitimate emails

- RQ1: We also checked the detection rate for legitimate emails and found no significant difference ($p = 0.681$).
- RQ2: We also checked the detection rate for legitimate emails and found no significant difference ($p = 0.259$).
- RQ3: We also checked the detection rate for legitimate emails and found no significant difference ($p = 0.7992$).

A.3. Results for non-video retention

To control for an effect on the general population, we also tested the difference between the initial phase and the retention phase for those that did not watch the video. We found no significant difference within this group ($t = 0.27085$, $df = 19$, $p = 0.7894$).

Appendix B. Interactive exercise

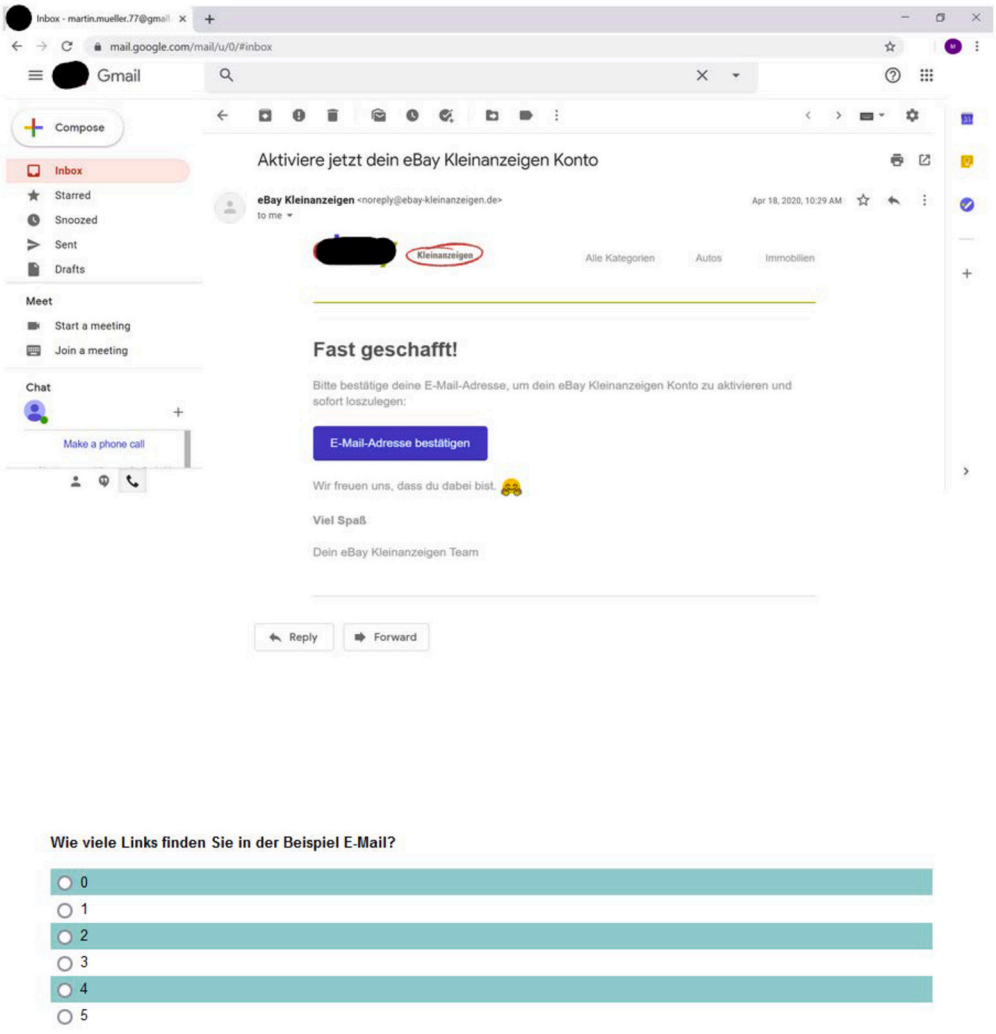

Fig. B.6. The Interactive Exercise that every participant had to succeed to proceed. We have censored the logos due to copyright laws. Translation: How many links can you find in the sample email?

Appendix C. Email example

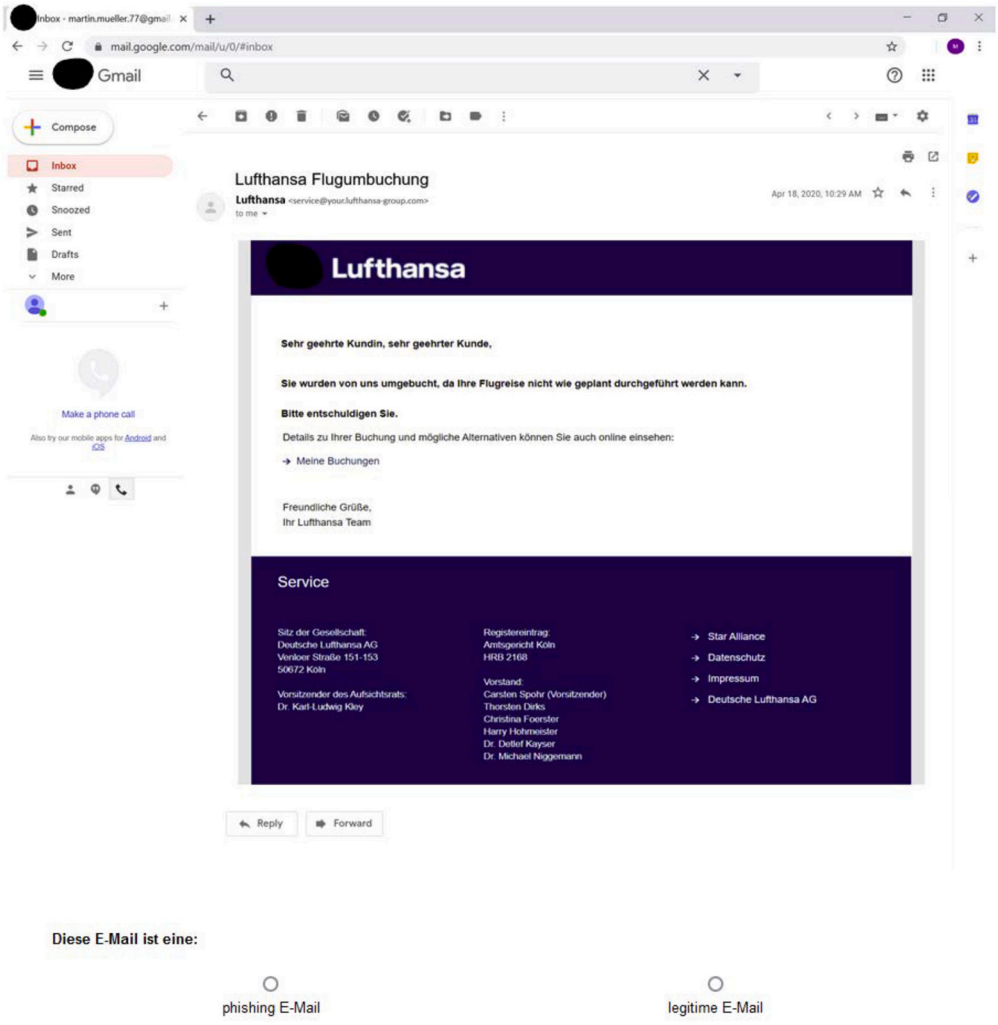

Fig. C.7. One example of an email that participants had to judge. We have censored the logos due to copyright laws. Translation: This email is a: Phishing email // Legitimate email.

Appendix D. Feedback question

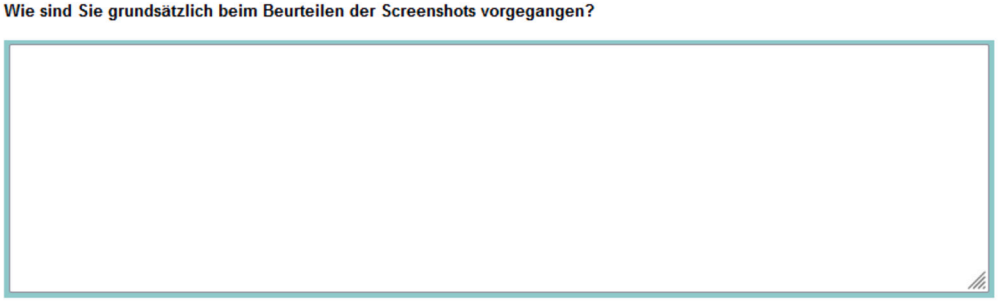

Fig. D.8. Feedback question on the participants' approach to their decision. Translation: How did you basically go about judging the screenshots? [free text].

Appendix E. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cose.2023.103620.

References

Abawajy, J., 2014. User preference of cyber security awareness delivery methods. Behav. Inf. Technol. 33, 237–248. https://doi.org/10.1080/0144929X.2012.708787.

Albakry, S., Vaniea, K., Wolters, M.K., 2020. What is this URL's destination? Empirical evaluation of users' URL reading. In: Conference on Human Factors in Computing Systems. ACM, New York, NY, US, pp. 1–12.

Althobaiti, K., Meng, N., Vaniea, K., 2021. I don't need an expert! Making URL phishing features human comprehensible. In: Conference on Human Factors in Computing Systems. ACM, New York, NY, US, pp. 1–17.

Anti-Phishing Working Group, 2022. Phishing activity trends report. Technical Report. Anti-Phishing Working Group. https://docs.apwg.org/reports/apwg_trends_report_q3_2022.pdf.

Arachchilage, N.A.G., Flechais, I., Beznosov, K., 2014. Poster: a game storyboard design for avoiding phishing attacks. In: Tenth Symposium on Usable Privacy and Security. USENIX, Berkeley, CA, US, pp. 1–2. https://cups.cs.cmu.edu/soups/2014/posters/soups2014_posters-paper39.pdf.

Arachchilage, N.A.G., Love, S., Beznosov, K., 2016. Phishing threat avoidance behaviour: an empirical investigation. Comput. Hum. Behav. 60, 185–197. https://doi.org/10.1016/j.chb.2016.02.065.

Berens, B.M., Dimitrova, K., Mossano, M., Volkamer, M., 2022. Phishing awareness and education – when to best remind? In: Symposium on Usable Security and Privacy. Internet Society, Reston, VA, US, pp. 1–15.

Beyer, M., Ahmed, S., Doerlemann, K., Arnell, S., Parkin, S., Sasse, A., Passingham, N., 2015. Awareness is only the first step: a framework for progressive engagement of staff in cyber security. Business white paper: Hewlett Packard. https://www.riscs.org.uk/wp-content/uploads/2015/12/Awareness-is-Only-the-First-Step.pdf.

Canfield, C., Fischhoff, B., Davis, A., 2015. Poster: using signal detection theory to measure phishing detection ability and behavior. In: Eleventh Symposium on Usable Privacy and Security. USENIX, Berkeley, CA, US, pp. 1–2. https://cups.cs.cmu.edu/soups/2015/posters/soups2015_posters-final13.pdf.

Canova, G., Volkamer, M., Bergmann, C., Berens, B., 2015a. NoPhish app evaluation: lab and retention study. In: Workshop on Usable Security. Internet Society, Reston, VA, US, pp. 1–10.

Canova, G., Volkamer, M., Bergmann, C., Borza, R., Reinheimer, B., Stockhardt, S., Tenberg, R., 2015b. Learn to spot phishing URLs with the Android NoPhish App. In: Ninth World Conference on Information Security Education. Springer, Cham, pp. 87–100.

Chang, L.Y., Coppel, N., 2020. Building cyber security awareness in a developing country: lessons from Myanmar. Comput. Secur. 97, 101959. https://doi.org/10.1016/j.cose.2020.101959.

Cohen, J., 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol. Bull. 70, 213–220.

Field, A., 2013. Discovering Statistics Using IBM SPSS Statistics. Sage Publications, Thousand Oaks, CA, US.

Franz, A., Zimmermann, V., Albrecht, G., Hartwig, K., Reuter, C., Benlian, A., Vogt, J., 2021. SoK: still plenty of phish in the sea — a taxonomy of user-oriented phishing interventions and avenues for future research. In: Seventeenth Symposium on Usable Privacy and Security. USENIX, Berkeley, CA, US, pp. 339–358. https://www.usenix.org/conference/soups2021/presentation/franz.

Garg, V., Camp, L.J., Mae, L., Connelly, K., 2011. Designing risk communication for older adults. In: Seventh Symposium on Usable Privacy and Security. USENIX, Berkeley, CA, US, pp. 20–22.

Gokul, C.J., Pandit, S., Vaddepalli, S., Tupsamudre, H., Banahatti, V., Lodha, S., 2018. PHISHY - a serious game to train enterprise users on phishing awareness. In: Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts. ACM, New York, NY, US, pp. 169–181.

Gonzalez, R., Locasto, M.E., 2015. An interdiscplinary study of phishing and spear-phishing attacks. In: Eleventh USENIX Conference on Usable Privacy and Security. http://cups.cs.cmu.edu/soups/2015/papers/eduGonzales.pdf.

Hamdani, K.J., Mustafa, M.I.E., 2021. Effectiveness of Online Anti-Phishing Delivery methods in raising Awareness among Internet Users. Master's thesis. Luleå University of Technology, Department of Computer Science, Electrical and Space Engineering.

Hart, S., Margheri, A., Paci, F., Sassone, V., 2020. Riskio: a serious game for cyber security awareness and education. Comput. Secur. 95, 101827. https://doi.org/10.1016/j.cose.2020.101827.

Jampen, D., Gür, G., Sutter, T., Tellenbach, B., 2020. Don't click: towards an effective anti-phishing training. A comparative literature review. Hum.-Cent. Comput. Inf. Sci. 10. https://doi.org/10.1186/s13673-020-00237-7.

Kirlappos, I., Sasse, M.A., 2012. Security education against phishing: a modest proposal for a major rethink. IEEE Secur. Priv. 10, 24–32. https://doi.org/10.1109/MSP.2011.179.

Kumaraguru, P., Cranshaw, J., Acquisti, A., Cranor, L., Hong, J., Blair, M.A., Pham, T., 2009. School of phish: a real-world evaluation of anti-phishing training. In: Fifth Symposium on Usable Privacy and Security. ACM, New York, NY, US, pp. 1–12.

Kunz, A., Volkamer, M., Stockhardt, S., Palberg, S., Lottermann, T., Piegert, E., 2016. Nophish: evaluation of a web application that teaches people being aware of phishing attacks. In: Informatik 2016 P-259, pp. 509–518.

Lastdrager, E., Gallardo, I.C., Hartel, P., Junger, M., 2017. How effective is anti-phishing training for children? In: Thirteenth Symposium on Usable Privacy and Security. USENIX, Berkeley, CA, US, pp. 229–239. https://www.usenix.org/conference/soups2017/technical-sessions/presentation/lastdrager.

Mayer, P., Volkamer, M., Kauer, M., 2014. Authentication schemes - comparison and effective password spaces. In: International Conference on Information Systems Security. Springer, Cham, pp. 204–225.

McHugh, M.L., 2012. Interrater reliability: the kappa statistic. Biochem. Med. 22, 276–282. https://hrcak.srce.hr/89395.

Misra, G., Arachchilage, N.A.G., Berkovsky, S., 2017. Phish phinder: a game design approach to enhance user confidence in mitigating phishing attacks. http://arxiv.org/abs/1710.06064. CoRR. arXiv:1710.06064, 2017.

Mossano, M., Berens, B., Heller, P., Beckmann, C., Aldag, L., Mayer, P., Volkamer, M., 2022. SMILE - smart eMail link domain extractor. In: Computer Security. ESORICS 2021 International Workshops. Springer, Cham, pp. 403–412.

Mossano, M., Vaniea, K., Aldag, L., Düzgün, R., Mayer, P., Volkamer, M., 2020. Analysis of publicly available anti-phishing webpages: contradicting information, lack of concrete advice and very narrow attack vector. In: European Symposium on Security and Privacy Workshops. IEEE, New York, NY, US, pp. 130–139.

Neumann, S., Reinheimer, B., Volkamer, M., 2017. Don't be deceived: the message might be fake. In: Trust, Privacy and Security in Digital Business. Springer, Cham, pp. 199–214.

Nisbett, R.E., Wilson, T.D., 1977. The halo effect: evidence for unconscious alteration of judgments. J. Pers. Soc. Psychol. 35, 250–256. https://doi.org/10.1037/0022-3514.35.4.250.

Onashoga, A.S., Ojo, O.E., Soyombo, O.O., 2019. Securix: a 3D game-based learning approach for phishing attack awareness. J. Cybersecurity Technol. 3, 108–124. https://doi.org/10.1080/23742917.2019.1624011.

Petelka, J., Zou, Y., Schaub, F., 2019. Put your warning where your link is: improving and evaluating email phishing warnings. In: 2019 Conference on Human Factors in Computing Systems. ACM, New York, NY, US, pp. 1–15.

Pienta, D., Thatcher, J.B., Johnston, A.C., 2018. A taxonomy of phishing: attack types spanning economic, temporal, breadth, and target boundaries. In: Workshop on Information Security and Privacy. Association for Information Systems, Atlanta, GA, US, pp. 1–18. https://aisel.aisnet.org/wisp2018/19.

Reinheimer, B., Aldag, L., Mayer, P., Mossano, M., Duezguen, R., Lofthouse, B., von Landesberger, T., Volkamer, M., 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. USENIX, Berkeley, CA, US, pp. 259–284. https://www.usenix.org/conference/soups2020/presentation/reinheimer.

Reynolds, J., Kumar, D., Ma, Z., Subramanian, R., Wu, M., Shelton, M., Mason, J., Stark, E., Bailey, M., 2020. Measuring identity confusion with uniform resource locators. In: Conference on Human Factors in Computing Systems. ACM, New York, NY, US, pp. 1–12.

Sasse, A., Hielscher, J., Friedauer, J., Peiffer, M., Menges, U., 2022. Warum IT-Sicherheit in Organisationen einen Neustart braucht. In: BSI - 18. Deutscher IT-Sicherheitskongress 2022. BSI, Berlin, DE, pp. 1–15. https://www.researchgate.net/publication/358277373_Warum_IT-Sicherheit_in_Organisationen_eien_Neustart_braucht.

Sheng, S., Holbrook, M., Kumaraguru, P., Cranor, L.F., Downs, J., 2010. Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions. In: Conference on Human Factors in Computing Systems. ACM, New York, NY, US, pp. 373–382.

Singhal, M., Levine, D., 2019. Analysis and categorization of drive-by download malware. In: Fourth International Conference on Computing, Communications and Security. IEEE, New York, NY, US, pp. 1–4.

Sood, A.K., Zeadally, S., 2016. Drive-by download attacks: a comparative study. IT Prof. 18, 18–25. https://doi.org/10.1109/MITP.2016.85.

Tschakert, K.F., Ngamsuriyaroj, S., 2019. Effectiveness of and user preferences for security awareness training methodologies. Heliyon 5. https://doi.org/10.1016/j.heliyon.2019.e02010.

Volkamer, M., Renaud, K., Reinheimer, B., 2016. TORPEDO: TOoltip-poweRed phishing email DetectiOn. In: Thirtyfirst ICT Systems Security and Privacy Protection. Springer, Cham, pp. 161–175.

Volkamer, M., Renaud, K., Reinheimer, B., Rack, P., Ghiglieri, M., Mayer, P., Kunz, A., Gerber, N., 2018. Developing and evaluating a five minute phishing awareness video. In: Trust, Privacy and Security in Digital Business. Springer, Cham, pp. 119–134.

Votipka, D., Abrokwa, D., Mazurek, M.L., 2020. Building and validating a scale for secure software development self-efficacy. In: Conference on Human Factors in Computing Systems. ACM, New York, NY, US, pp. 1–20.

Wang, J., Li, Y., Rao, H.R., 2016. Overconfidence in phishing email detection. J. Assoc. Inf. Syst. 17. https://doi.org/10.17705/1jais.00442.

Wash, R., Cooper, M.M., 2018. Who provides phishing training? Facts, stories, and people like me. In: 2018 CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, US, pp. 1–12.

Wen, Z.A., Lin, Z., Chen, R., Andersen, E., 2019. What.Hack: engaging anti-phishing training through a role-playing phishing simulation game. In: Conference on Human Factors in Computing Systems. ACM, New York, NY, US, pp. 1–12.

Zielinska, O.A., Tembe, R., Hong, K.W., Ge, X., Murphy-Hill, E., Mayhorn, C.B., 2014. One phish, two phish, how to avoid the Internet phish: analysis of training strategies to detect phishing emails. In: Human Factors and Ergonomics Society Annual Meeting, vol. 58, pp. 1466–1470.

Benjamin Berens - Research assistant in the SECUSO research group of Prof. Dr. Melanie Volkamer at the Karlsruhe Institute of Technology since May 2017. He finished his Business Psychology Bachelor program at the University of Applied Science Fresenius

and graduated from the Technische Universität Darmstadt with a master's degree in psychology. Prof. Dr Melanie Volkamer supervised his master thesis on the evaluation and further development of the TORPEDO thunderbird Add-On. The overall goal of his work is to look at the status quo of security policies, awareness-raising and training activities and develop improved new strategies and tools based on the newest research.

Mattia Mossano - Research assistant in the SECUSO research group of Prof. Dr. Melanie Volkamer at the Karlsruhe Institute of Technology since December 2019. Before joining KIT, he completed his master's degree in cognitive science at the School of Informatics of the University of Edinburgh. The thesis dealt with the investigation of general advice against phishing attacks found on various public websites. He also wrote a master's thesis in philosophy at the University of Genoa, which criticized the use of evolutionary algorithms to generate general AIs. His main research interests are anti-phishing awareness, security support tools and accessible cybersecurity.

Prof. Dr. Melanie Volkamer - Full professor of Security Engineering at the Karlsruhe Institute of Technology and head of the SECUSO research group since 2011. From August 2016 until March 2018 she was a Professor (Kooperationsprofessur) at the Department of Computer Science of Technische Universität Darmstadt. From December 2015 until December 2018, she has been appointed Full Professor for Usable Privacy and Security at Karlstad University. Before, she was an Assistant Professor at TU Darmstadt. From May 1st to August 31st 2011, she worked as a visiting researcher at CMU/CUPS. She has made contributions in the fields of web security and privacy, electronic voting, authentication and mobile security and privacy.