Checking, nudging or scoring? Evaluating e-mail user security tools

Sarah Y. Zheng and Ingolf Becker, *UCL*

https://www.usenix.org/conference/soups2023/presentation/zheng

This paper is included in the Proceedings of the Nineteenth Symposium on Usable Privacy and Security.

August 7–8, 2023 • Anaheim, CA, USA

978-1-939133-36-6

Open access to the Proceedings of the Nineteenth Symposium on Usable Privacy and Security is sponsored by USENIX.

Checking, nudging or scoring? Evaluating e-mail user security tools

Sarah Y. Zheng
UCL

Ingolf Becker
UCL

Abstract

Phishing e-mail threats are increasing in sophistication. Technical measures alone do not fully prevent users from falling for them and common e-mail interfaces provide little support for users to check an e-mail's legitimacy. We designed three e-mail user security tools to improve phishing detection within a common e-mail interface and provide a formative evaluation of the usability of these features: two psychological nudges to alert users of suspicious e-mails and a "check" button to enable users to verify an email's legitimacy. Professional e-mail users ($N = 27$) found the "suspicion score" nudge and "check" button the most useful. These alerted users of suspicious e-mails, without harming their productivity, and helped users assert trust in legitimate ones. The other nudge was too easily ignored or too disruptive to be effective. We also found that users arrive at erroneous judgements due to differing interpretations of e-mail details, even though two-thirds of them completed cybersecurity training before. These findings show that usable and therefore effective e-mail user security tools can be developed by leveraging cues of legitimacy that augment existing user behaviour, instead of emphasising technical security training.

1 Introduction

E-mail has been one of the most pervasive forms of digital communication since the introduction of the internet. So much so, that the medium remains an attractive threat vector for adversaries to exploit [50]. Phishing e-mails, in which impersonated sources typically seek to gain money or sensitive data from a target recipient, have caused major security breaches, financial losses and psychological damage to unsuspecting users, making it a lucrative business for organised crime [15, 23, 33, 61].

Technical detection systems may capture the majority of phishing attacks, but do not fully prevent users from falling for them. As users are commonly regarded as the "last line of defence" [2], organisations invest in cybersecurity education for their employees and inform the public of potential scams. However, anti-phishing education and publicly available anti-phishing advice may not be as effective as hoped for [13, 38, 47, 57, 59].

An alternative way to help users disengage with suspicious e-mails is to enhance common e-mail interfaces to equip users with "just in time" decision-making tools. For instance, by nudging users to check sender information [49] or interactively showing the trustworthiness of URLs found in e-mails [54, 77]. Such developments showed promising results to decrease phishing susceptibility, but have been sparse and require further exploration [27].

Here, we provide an implementation-focused formative evaluation [71] of the usability of novel user-centric e-mail security concepts to help users detect suspicious e-mails in a common e-mail user interface (UI). First, we conceptualise (i) a "check" button to highlight indicators to help people assess both trustworthy and phishing e-mails, and (ii) a "collegiate phishing report" nudge and (iii) "suspicion score" nudge to make people aware of the possibility of phishing. We then examine how these security tool concepts affect users' e-mail processing behaviour, by collecting "think aloud" responses from professional e-mail users from one organisation (N=27) that processed e-mails in simulated Outlook e-mail interfaces without and then with the tools. Tool designs were updated following consistent feedback from at least five users over four iterations.

We find that the suspicion score nudge and final check button version were rated the most useful, and that people largely process the same pieces of e-mail information, but reason differently about them. This was surprising, as 18 of

USENIX Symposium on Usable Privacy and Security (SOUPS) 2023.
August 6–8, 2023, Anaheim, CA, United States.

the participants recalled completing at least one mandatory cybersecurity training before and 19 have a technical study background. This implies that worse detection is not necessarily due to negligence of relevant security indicators [90], but a lack of consensus on how to interpret online information. For example, whether e-mails sent from free e-mail providers should be suspected in a professional context. This resulted in the following contributions:

1. We identify and discuss three fundamental trade-offs that can guide further development of usable e-mail security tools: (i) highlighting cues of desired (i.e., legitimate) vs. undesired (i.e., phishing) communication, (ii) enhancing users' existing behaviour vs. technical knowledge and (iii) not harming productivity for security.

2. We open-sourced our methods and data via GitHub and Open Science Framework (OSF) to encourage more studies on e-mail user security tools. This includes the simulated Outlook UI and two adapted e-mail sets to fit participants' organisational context to closely mimic an e-mail processing experience.

2 Related work

2.1 User-centric security interventions

People are thought to be bad at detecting phishing e-mails due to a lack of cybersecurity knowledge or awareness [5, 7, 38, 80, 83] and incautious e-mail processing behaviour [22, 31, 36, 44, 46, 76, 90]. The majority of interventions to improve human phishing detection thus focused on developing training and education programs [27]. Examples range from conventional education materials [6, 13, 16, 34, 36, 68, 82, 88] and serious games [9, 19, 28, 30, 40, 41, 69, 84], to phishing simulations [8, 18, 39, 40, 78]. While they increase user awareness of phishing threats, these programs require more frequent engagement than often is the case to stay effective in the long term [13, 59]. This waning effect may be due to the timing of educational programs before or after, but not during critical decision-making moments [27].

A different stream of user-centric security interventions aimed to make people aware of security-relevant information *during* decision-making. Earlier works used browser-based security warnings to prevent users from browsing suspicious domains [4, 26, 58, 67, 87]. Although such warnings are moderately helpful, they are prone to warning fatigue [4]. Alternatively, digital signatures may be used to add trust signals to e-mails [86], although criminals typically adopt them too once their use becomes widespread, as happened with SSL certificates [1].

Others have highlighted dubious domains [43] or e-mail sender information [49], but found limited detection improvements. This is likely due to users' misinterpretation of URLs [5]. A more promising approach provided users with interactive URL reports when they engaged with links found in e-mail messages [7, 54, 77]. By informing users about why a URL may be suspicious, these tools provide both educational and awareness-raising value. These works underline the potential of user-centric security interventions embedded in e-mail interfaces.

We expanded on this idea of aiding users during decision-making in an e-mail UI [27, 90]. Specifically, we used two under-explored concepts in e-mail security to design novel security features: 1. **enhancing users' confidence in trusting legitimate e-mails**, instead of following the predominant paradigm of enhancing phishing detection, and 2. **psychological nudges**, where slight changes in a UI improve decision-making, without forcing users to engage with those changes [72]. For instance, participants in a simulated e-commerce purchase displayed more secure behaviour when they received a notification that emphasised how they can cope with online shopping risks [73]. A study on phishing detection used a social alert in suspicious e-mails, saying that a high percentage of colleagues received the same e-mail [49]. Even though these works found small detection improvements, these findings suggest that nudging users with short and directly applicable information embedded in task systems can improve security behaviours. The potential of e-mail security nudges is discussed further in Franz et al. [27].

In line with the two concepts, we devised a "check button" to help users asses any e-mail's legitimacy, and two different nudges to alert users of suspicious e-mails. The first nudge contained a social cue, similar to Nicholson, Coventry, and Briggs [49], and also explained users what suspicious signs to look out for. The second nudge alerted users of potentially suspicious e-mails and showed recommended actions. With these features, we aimed to improve phishing detection by better supporting how users reason about e-mails while they are processing them.

2.2 Processing e-mail for communication versus security

E-mail processing has been described as comprised of "primary" and "secondary" tasks in the cybersecurity context [64]. The primary task refers to the main function of e-mail, i.e., communicating with others through digital means, which involves scanning, prioritising and responding to e-mail messages. The secondary task is the security check to decide if an e-mail is in fact legitimate and responding accordingly. On the one hand, we cannot reasonably expect users to focus on the secondary task [64]. On the other hand, even if we do, making the secondary task the main focus will inflate users' suspicions [53, 66, 70]. It is therefore a vital research challenge to design user-centric security interventions that augment users' secondary e-mail processing task, without harming their primary task.

The first step towards this goal is a deep understanding of

how users switch between the primary and secondary task in real-world contexts. Various studies have characterised aspects of how users process e-mails and the usability of adapted e-mail interfaces [11, 12, 21, 29, 45, 51, 79], but these provide limited or no insight into how users switch to the secondary task. The few qualitative works that do focus on how users reason about phishing find that both experts' and non-experts' e-mail processing involve understanding the e-mail context, finding surprising elements that lead to suspicion and acting on that suspicion [81, 83]. However, as these studies used phishing detection as the primary task [81] and relied on respondents who remembered a previously received phishing e-mail [83], it is unclear to what extent these results generalise to real-life contexts.

Thus, to engage users with the secondary task when they should suspect an e-mail, we need to understand how users change their reasoning from the primary to the secondary task. Then we can see how our proposed security features affect users' processing behaviour. Hence, we first analyse how users process e-mails without any security interventions, and then how our designs affect this behaviour.

3 Methods

Our formative evaluation [71] focuses on understanding what drives (un)usability of our novel e-mail user security tool concepts. To this end, we used qualitative methods to obtain an in-depth understanding of how our e-mail security tool designs affected users' e-mail processing behaviour and iterative design to make small short-term adjustments according to consistent user feedback. In this section we describe the study setup, the principled approach to our designs, our participants sample and analysis.

3.1 Participants

Twenty-seven participants performed the in-person e-mail processing task. They were recruited through e-mail invitations sent to staff at the researchers' institute. We only recruited staff from our institute, because (i) all staff were known to be experienced e-mail users, (ii) they could come to the session in-person, (iii) they would be familiar with the presented task context (e.g. e-mails from the same institute), (iv) they are likely to be used to the Outlook e-mail client, as their professional e-mails are processed through Outlook, and (v) they represent a working office population that relies substantially on e-mail communication. All were compensated with a £20 Amazon voucher. Their roles ranged from support staff to lecturers. The study was approved by our departmental Ethics Committee.

3.2 Task

To understand how users process e-mails, we asked participants to reason out loud while processing e-mails in simulated inboxes. They were told the study aimed to gather feedback on the usability of new e-mail interfaces and not security tools, to avoid biased responses [52]. We created a basic Outlook e-mail interface as shown in Figure 1, to which we added our security tool designs. Each participant had an in-person session of 45–60 minutes with the main researcher who sat down next to them.

After welcoming the participant and obtaining their informed consent, the researcher started an anonymous audio recording of the session. Participants first answered questions on their general e-mail use and were then instructed about the main task. They had to process e-mails as if they were professor Alex Carter in health informatics and talk through what they were doing and why. Participants were never told about phishing detection before or during the task. Only the security tool designs in the task could have prompted them to look out for phishing e-mails, as intended.

Each participant interacted with four different inboxes, one after another. They always started with the "control" inbox, i.e., without any new tools (Figure 1). The next three inboxes each contained one of the three security tools (see Section 3.3 and Figure 2) in random order. Each inbox contained eight or nine e-mails based on e-mails previously received by colleagues at the same institute to provide a familiar context (total $N_{legitimate} = 33$), and one or two phishing e-mails adapted from those previously received by academic institutes (total $N_{phishing} = 6$). Two phishing e-mails contained a malicious URL, purporting to be a Zoom meeting invite and Microsoft password reset. Two spearphishing e-mails seemed to come from professor colleagues requesting an urgent action. One phishing e-mail presented a fake paid mentor program, one "Nigerian prince"-style scam, and (see details on GitHub). Nearly all e-mails were made to directly address professor Alex Carter. Each e-mail could be replied to, forwarded, deleted, archived or moved to "junk". When participants wanted to reply to an e-mail, a text editor appeared through which they typed their reply and "sent" the message. We deemed these functionalities sufficient to simulate the experience of processing e-mails for human end users, as they cover the majority of user actions to process e-mails. This was confirmed by a participant's remark *"this feels like going back to work"* when they started the task.

To facilitate users with out-loud reasoning, the researcher asked participants to explain what they were looking at, to elaborate why they responded in certain ways to the e-mails, and, when participants explicitly mentioned that an e-mail looked legitimate or suspicious, why they thought so. The researcher also took written notes of any significant observations and asked if participants noticed any new feature when they did not interact with them during the first 2–3 minutes

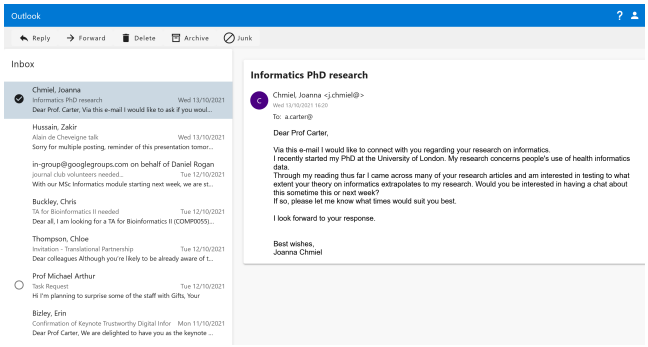


Figure 1: **Screenshot of an example e-mail in the "control" inbox.** The interface mimics the Outlook web client. Our security feature designs were added to this basic UI.

of viewing the inbox, to see why they had not (see study protocol on OSF). When doing so, we were careful not to mention any security concept, nor asking them to use the tool, to avoid biasing participants' processing behaviour. This way, any increased security awareness was merely the result of participants noticing the new security tool.

Every seven minutes, the next inbox automatically appeared, until participants saw all four inboxes. We kept the time spent on each inbox constant across participants to rule out the possibility that user engagement changed as a result of different times spent with a particular tool and ensure a study duration proportional to participants' compensation. We were aware that doing so traded off measuring detection accuracy for a reproducible qualitative method to evaluate usability. Efficacy will accordingly be described in terms of qualitative observation, not statistical comparison.

After completing the main task, participants gave feedback, voted which tool they found most useful and would use in real life, how many phishing e-mails they receive themselves, how much cybersecurity training they completed before and answered demographic questions (age, gender, education level, study background).

3.3 Rationale for tool designs

We designed our security features with two goals in mind: to help human users detect phishing e-mails and to assure users of the legitimacy of genuine e-mails. We took a principled approach. All designs had to be (i) user-centric, i.e., keeping humans "in the loop", (ii) accessible, i.e., easy to understand and use, and (iii) available at all times, which implies embedding new functionalities within the e-mail UI. The latter departs from conventional cybersecurity training, which may align with principles (i) and (ii), but not (iii). We defined three tool concepts: (i) "check" button, (ii) collegiate phishing report nudge, (iii) suspicion score nudge. We also kept the designs relatively small, in the form of an inbox add-on.

As part of our formative evaluation, we made small adjustments to the tool designs after at least five users gave us the same feedback. This resulted in four user-driven design iterations. The "check" button was updated three times, the "collegiate phishing report" nudge updated twice and the "suspicion score" nudge once. Figure 2 shows the final versions of the three tool designs, Appendix A depicts each iteration. Note that all updates were display-related changes and did not change key functionalities.

3.3.1 Check button

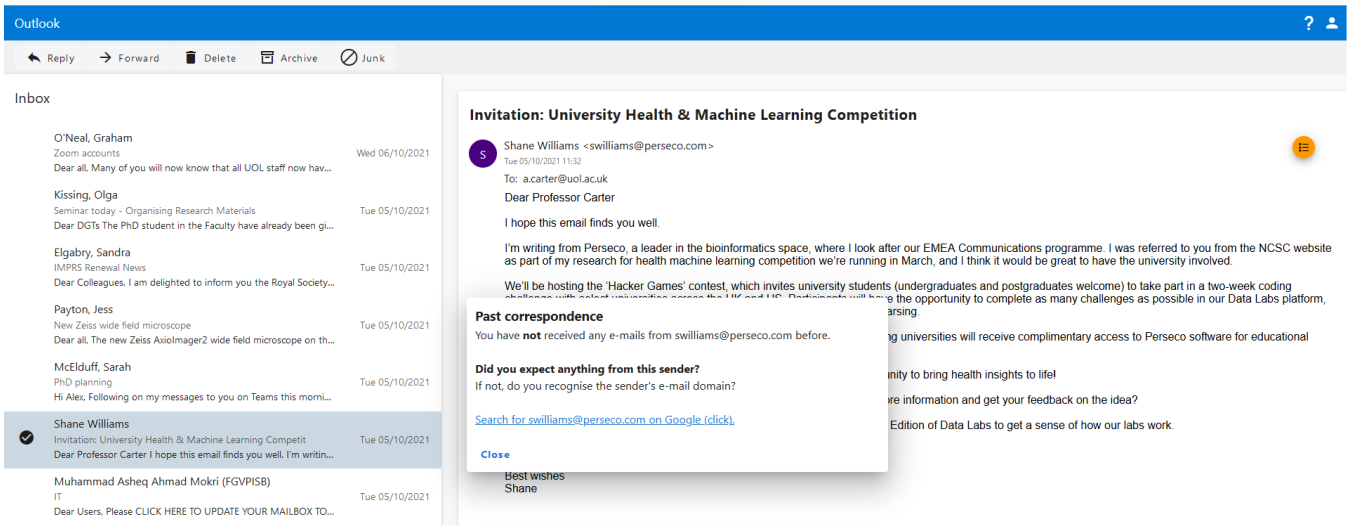
The first version of the "check" button sat in the task ribbon next to the "Junk" button. It aimed to provide users information on whether to trust a selected e-mail, based on common heuristics used by IT experts [81]. It could display overviews of (i) a dissection of the true URLs of any links found in the selected e-mail, since users often misinterpret URLs [5], (ii) the sender's name and e-mail address with short pieces of advice on what to do in case of mismatches in said details, and (iii) past e-mails received from the sender e-mail address. We expected these simple "checks" to help users when they are unsure if they could trust an e-mail by showing how URLs and sender details should be parsed and conjugated, as previous work implied that many users lack such reasoning [90].

Following consistent user observations, only the "past correspondence" check was kept in the last iteration. We placed the button closer to the e-mail sender details to which its functionality applied, following previous security feature design recommendations [74, 75], and simplified the button. If the user received e-mails from the sender's e-mail before, they are shown in a list with the date, time and subject line. If no past correspondence exists, the check information asks the user if they expected anything from the sender. If they did not, they are asked to double check the sender's e-mail domain. See Figure 2A and Appendix Figure 3.

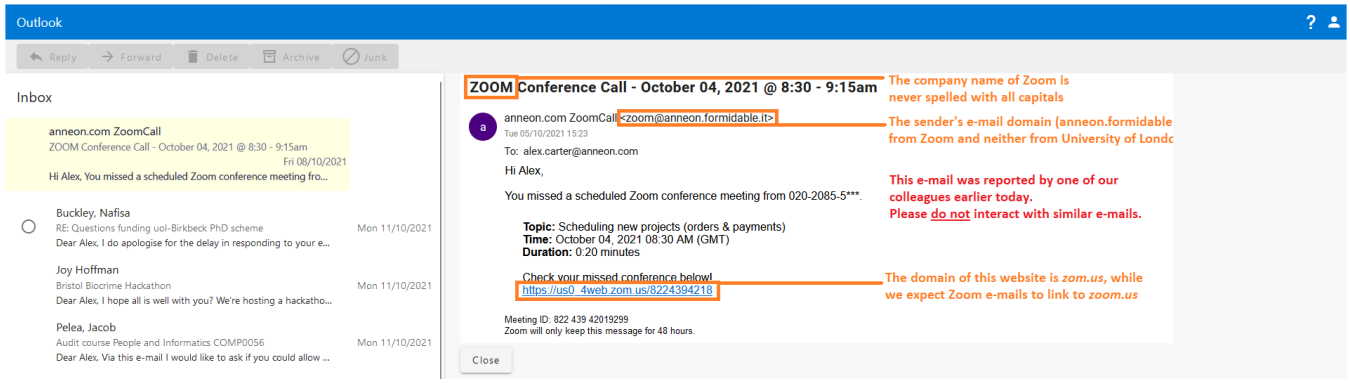
3.3.2 Nudge 1: Collegiate phishing report

The "collegiate phishing report" aimed to shift users' cognitive frame to investigating e-mail legitimacy [27], by using a socially oriented nudge that read "This e-mail was reported as suspicious today by one of our colleagues". The first version displayed this text in an orange warning banner between the Outlook task ribbon and e-mails display. When users clicked on it, a floating display appeared on top of the inbox with a screenshot of the phishing e-mail that was purportedly reported by a colleague, with annotations of all the suspicious cues in the e-mail that users had to look out for, and a general recommendation to not interact with similar e-mails. In the last iteration, the nudge looked like a new e-mail at the top of the e-mails list to increase user engagement. When users clicked on it, they saw the nudge text in the e-mail display with the fully annotated e-mail message and action recom-

A. Check button



B. Nudge 1: Collegiate phishing report



C. Nudge 2: Suspicion score

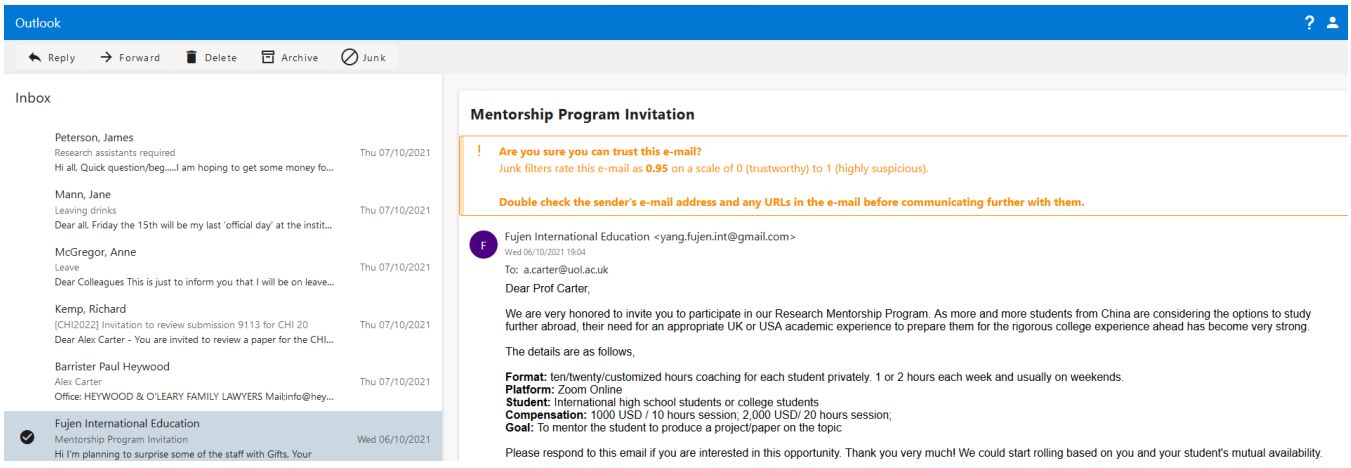


Figure 2: **Final designs of each security feature: A.** the past correspondence check button, **B.** collegiate phishing report nudge, **C.** the suspicion score nudge.

Iteration	Check button	Nudge 1: Collegiate phishing report	Nudge 2: Suspicion score
1 ($N = 8$)	The majority of users were unaware of the button until nudged towards it (after 2–3 minutes); users did not explore all sub menu items	Users tended not to click on the warning banner or got confused about which e-mail the warning is referring to	Users did not read all provided information, but found the orange colour positively alerting and useful
2 ($N = 7$)	Users remained unaware of the button until the researcher pointed it out, but also often did not see the benefit of the provided information.	Users did not like pop-up windows and often felt urged to close it right away	(design did not change)
3 ($N = 5$)	Users remained unaware of the button until the researcher pointed it out; the 'past correspondence' element was deemed useful	(design did not change)	(design did not change)
4 ($N = 7$)	Most users who noticed and started using the button found it very useful	More users skimmed over the warning content, some users found this and the suspicion score generally useful as they alerted them of suspicious e-mails	Users did not read all provided information, but found the orange colour positively alerting and useful; subtle text formatting edits did not lead to significantly more users applying the recommended actions

Table 1: Summary of user feedback on the security tool designs in each iteration

mendations. These annotations could not be removed and users could remove the nudge with the close button below it. See Figure 2B and Appendix Figure 4.

3.3.3 Nudge 2: Suspicion score

"Suspicion score" nudges were added to the phishing e-mails to prompt users to make a conscious effort to assess their legitimacy. This nudge was displayed in an orange warning banner between the task ribbon and e-mail display, and said "Are you sure you can trust this e-mail?". Below this line, it showed how suspicious the e-mail was on a scale from 0 to 1 and immediate action recommendations to nudge users' coping ability [73]. We chose to describe the score and scale to encourage users to think of potential false positives, e.g. when an e-mail has a lower suspicion score around 0.60, and thus take the recommended actions. Contrary to some inboxes that subtly add "(SPAM?)" in plain text to junk e-mail subject lines, we expected that our approach on e-mails in the main inbox would have a greater effect on users' vigilance and that explaining why the nudge was displayed would address user concerns over e-mails that could unwantedly be marked as junk. The amount of information was reduced and the text formatting changed slightly for the last iteration. See Figure 2C and Appendix Figure 5.

3.4 Thematic analysis

We used thematic analysis (TA) of a reflexive nature [17] to understand e-mail users' motives and considerations while they perform a typical e-mail processing task and evaluate how our tools affected these as a measure of usability. Given limited prior works on the qualitative relation between users' real-life e-mail processing behaviour and security, we interpreted the data inductively—without prior theories or hypotheses. After transcribing all session recordings, two researchers independently annotated one transcript, discussed the annotation codes and re-coded the same transcript to agree on the granularity of the coding approach. The first author, who conducted all participant sessions, annotated all remaining transcripts and freely added new codes to document all their user observations, in line with an interpretivist stance [17]. After completing all annotations, we iteratively extracted and refined themes through further discussions driven by the data. All transcripts and the full code book are available via OSF. We do not report inter-rater agreement scores, as they are inappropriate in reflexive TA [17].

4 Results

We performed an implementation-focused formative evaluation [71] of the usability of novel e-mail user security tool concepts based on highlighting legitimacy instead of phishing cues and nudging. We implemented the tool designs in simulated Outlook inboxes and let 27 professional e-mail users (mean age = 33.2 (SD = 7.2); 48% male; mean number of e-mail accounts = 4 (SD = 2.1); 19 with a Science, Technology, Engineering or Maths background) process e-mails in them while reasoning out loud. Their feedback was used as ongoing input for small short-term tool adjustments after at least five users provided similar feedback, resulting in four design iterations (Table 1).

Through thematic analysis, we gained a deep qualitative understanding of how our tools affected users' e-mail processing. Specifically, to understand why certain tools were (un)usable, it was key to understand how users reason without and then with our tools. A total of nine top-level codes and 86 secondary-level codes describe all user observations (see Appendix B and full code book on OSF). Based on these codes, we uncovered four overarching themes that capture how users reason about e-mails without our security tools (Section 4.1), and four themes that reflect how our security features affected them (Section 4.2).

4.1 Users' e-mail processing behaviour

The final codes naturally evolved to mirror users' primary and secondary e-mail processing behaviour. The "processing reasons" codes capture users' primary e-mail processing, and codes under "signals suspicious" and "signals non-suspicious" capture what users consider when they explicitly judge e-mails to be suspicious or non-suspicious (i.e., secondary e-mail processing). "Intended processing actions" reflect what users do with e-mails that do not raise any suspicion (i.e., primary processing) and "mitigation strategies" reflect how users assess and manage suspicious e-mails (i.e., secondary processing). Similar codes under "processing reasons" and "signals (non-)suspicious" suggest that users can arrive at opposing conclusions and perform different actions based on the same reasons (Section 4.1.3). Together with "prioritisation approach" and "prior experiences", these codes gave rise to the following four themes that describe how users generally reason about e-mails.

4.1.1 Content relevance

Most users first considered the relevance of the e-mail message content by judging the intent of the e-mail sender, before deciding what action to take. They either skimmed over subject lines, skimmed over the e-mail or read e-mails line by line right away. The importance of content relevance judgements is also reflected by the amount of "processing reasons" codes that relate to message contents ("high frequency", "important or urgent", "keep for reference", "not right audience or not personally targeted", "of personal interest", "outdated", "thread", "uninteresting or irrelevant"). These observations imply that users empathised with the e-mail task context.

The most common reason for users to find an e-mail suspicious was also based on e-mail message content. Out of all codes under "signals suspicious", "unexpected or funny content" has by far the most references. That is, most users seemed to assume legitimacy until they encountered e-mail content they perceived as odd. This accords with prior studies [52, 81], as well as psychological theory that people merely suspect things that are unlikely [24]. Further content-based reasons for users to be suspicious of an e-mail were "funny URLs", "requesting personal details", "urgent matter" and "fear appeal".

4.1.2 Relation to sender

Next, most users inferred their relationship with the perceived e-mail sender, by reading the sender's display name, a signature in the e-mail message and/or the actual sender's e-mail address. They made an assumption of how close they are to this sender according to the way the e-mail was written and the sender's e-mail domain. The "processing reasons" that reflect this theme are "unknown sender", "assume known or trusted sender", "from internal organisation", "automated e-mail", "newsletter". For example, user O112 assumed that the sender of a spear-phishing e-mail was indeed from the purported colleague professor: *"[. . .] a task request and it's from a professor and probably someone [who is] also a colleague of mine. And it's more personal because it starts with 'hi'."* This user assumed so, given the e-mail's informal writing style ("assume known or trusted sender"). To them, an "urgent request" may be reasonable to receive from a colleague and would not trigger suspicion.

Perceived closeness to e-mail senders was also the second most referenced factor that drove secondary processing. This is shown through "signals non-suspicious" codes "past correspondence", "internal e-mail", "trusted sender e-mail address" and "signals suspicious" codes "external sender", "non-professional sender e-mail", "no online info about sender organisation" and "unexpected sender or recipient name or e-mail address". An example of the latter, O11: *"This is a bit of a stranger and maybe someone genuinely called Olga Kissing. That is a bit suspect. So that depends on whether I actually knew that person. So I would just be suspicious from there."* The user reasoned that the sender's name sounded funny and therefore was untrustworthy, without reading the actual e-mail content. Only if they knew someone with that name, they might trust it.

4.1.3 Subjectivity in legitimacy perceptions

Users could have completely diverging assessments of the same e-mail, as the most attended to factors in e-mail processing described above are prone to subjective interpretation. For example, when users perceived an e-mail as not directed at them, we observed any of eight subsequent processing actions (see visualisations in Supplementary Materials on OSF). This aligns with prior work on user perceptions of "misdirected e-mail" [56]. We found bigger consensus among users that "unexpected or funny content" and "unexpected sender or recipient name or e-mail address", but not technical indicators, make an e-mail suspicious.

One scenario was that different users mentioned the same reason, but drew opposing conclusions for the very same e-mail. A prime example of this was in the case of a spear-

phishing e-mail sent from a GMail address purporting to be from a colleague professor. Out of all users who noticed this "unprofessional sender e-mail address", some users said it was junk straight away, whereas others responded without any suspicion. For example, user O19 commented: *"That sounds like a scam.[...] Because [of] the first bit. Also, it's from a GMail address."* First, they did not trust the e-mail because of the message content. They then noticed the sender's GMail domain, which further confirmed their suspicion.

Other users reasoned further about using private GMail addresses for work, e.g. user O14: *"If it's, like, a professor, same university. I would expect that communication would go in the same channel. So, like the University of London rather than a private e-mail. So I would maybe call that person and just, uh, ignore it, to be fair."* They would not expect a professional colleague to use a non-professional e-mail address to communicate with them. Their mitigation strategy would have been to call the sender to verify if they indeed sent the e-mail.

When viewing another e-mail from a GMail address, they described their suspicion of GMail accounts: *"And maybe they hate the Outlook interface by the university. [...] would not exclude it directly. That's the reason why I would look more on the content rather than, I mean, if it's like an e-mail, like an alpha numerical contact, like C H zero five, blah blah about to dot com, then I would think that not the right motivation is there."* They would consider both the e-mail message content and the sender's e-mail domain, but put more weight on the content. Similarly, user O26 appraised the e-mail content, but assumed that the same e-mail came from a known colleague, without mentioning any suspicion:

> *"Well, they've used a personal account, but they've signed it off as Professor Blackfield, and they work at the University of London [...] I would respond and say, sure, no problem. If it was someone I didn't know or the e-mail address was unfamiliar, I would probably ignore, delete. But in this instance [...] I presumed I know them. So I would say 'sure'."*

Another scenario was when users assessed different aspects of the same e-mail and drew opposing conclusions as a result. For example, O15 only looked at the message content of a phishing e-mail that indicated a missed Zoom conference and said: *"So this other e-mail is a Zoom conference call, but we missed it. If it is very important, [...] I would just mark it on my calendar to check this conference content or communicate with the people if necessary. But I would pin it if it is important."* They were not suspicious of the e-mail at all and overlooked the odd sender details and URL in the e-mail body. In contrast, user O27 first noticed unexpected sender details and marked the same e-mail as "junk": *"This e-mail address, Anneon.formidable, it is spam, so it's going to get junked. I do not do orders and payments. Somebody will tell me. I don't need an automated e-mail. So, that's junk."* It generally seemed that once any cue raised suspicion, users got rid of the e-mail as soon as possible or they started processing more information to substantiate their initial hunch—in line with findings from previous works [81, 83]. These examples show that within an e-mail, users assess different aspects, which leads to diverging legitimacy judgements.

4.1.4 User intents to UI functions

Our inbox simulation facilitated insights into how users translate their e-mail processing reasoning to how they interact with common inbox functionalities. We found that users had their own "mental models" of these functionalities. Strikingly, the vast majority of users deleted e-mails that they found suspicious, even though there was an option to mark e-mails as "Junk", e.g. O19: *"Junk it, delete it. Either way, get it out of the inbox. Like, I personally very rarely use junk to get rid of something."* This implies that most users do not distinguish between the type of "unwanted e-mails"—whether they thought the e-mails were uninteresting, irrelevant, or (potentially) malicious. They were usually treated the same way. It may thus not be practical for users to apply a different process to distinguish e-mails they found suspicious. User O13 even mentioned that they did not know that they could move e-mails to "Junk" themselves:

> *"**Researcher:** I also noticed that one of the e-mails that you thought was suspicious, you deleted it and you didn't say junk. Is that what you normally do as well?*
> ***User O13:** Yeah, I wouldn't necessarily say junk. That's not something we have, do we? [...] I never thought of to use that, possibly I'm ignorant. [...] I just, I never knew it existed, that we had a junk and we could put things in junk."*

Users also largely ignored the "Archive" button. E-mails that would be archived were usually deemed unnecessary and various users said they are unlikely to read archived e-mails ever again. As with suspicious e-mails, users may favour to delete anything irrelevant. A few users were concerned about e-mail storage and reasoned that deleting would therefore be better than archiving.

4.2 Effect of security features on users' e-mail processing

Thus far, we described users' overall reasoning while processing e-mails. Next, we examined how our security tools affected this processing behaviour. Through the iterative design process, we found that for tools to be usable, users value as little disruption to their primary task as possible and that simple changes such as the (symbolic) colour or placement of the tool matter more than content. As a result, the "suspicion

score" nudge received the most positive feedback in all iterations, with the "past correspondence" check from the "check" button as runner-up. In total, 9 participants found the "suspicion score" most useful, 7 participants the "check" button, 4 participants both the "suspicion score" and "check" button, 1 participant the "collegiate phishing report", 4 participants found all three designs most useful and 2 participants none of the tools.

All user interactions with and intra-task feedback on our tool designs were coded under "intervention feedback". When users adopted the given tool during the task and/or found it useful, it was coded as "positive". The other "intervention feedback" codes indicate points for improvement. Four themes emerged from these codes and "prior experiences", which together capture how users experienced the security features in relation to their e-mail processing. We summarised users' implicit (intra-task) and explicit (post-task) feedback in Table 1.

4.2.1 Usability of security information

Some "intervention feedback" pertained to the usability of information provided by the security features ("missing useful info", "functionality not clear", "too much information to process"). Fifteen users mentioned that the functionality of the "collegiate phishing report" nudge and/or the "check" button in the first three iterations was unclear, even though these tools were specifically designed to facilitate human interpretation of technical e-mail details. When users viewed the "check" button content in iterations 1–3, they often did not know how to interpret the provided information. For example, O12 in the first iteration:

> *"**Researcher:** When you see this, what are you thinking? [. . .]*
> ***User O12:** Whether there is malice or not. It's just not sufficiently. Well, it looks clunky [. . .] it would be much more useful to have something very clear saying 'this is safe, this is not safe' rather than giving me all this information."*

They were viewing a phishing e-mail and used the "check" button. After skimming over the provided check information, they pointed out that the information did not tell them whether to interpret the e-mail as suspicious or not. We expected users to be able to infer themselves whether they could trust an e-mail with the given details, but this did not seem the case. We considered displaying the information differently and adding more guidance in iteration 3, but they still found the amount of information too much, e.g. O16: *"Okay. Woah, so this confused me right away. So I just. Whatever. I just get out of here. Close. Because there's too much information. I don't understand anything. Lot of questions. A lot of, like, uh, sender details. And I don't know what they are."* We observed a similar negligence of technical information in the "suspicion score" nudge, which contained far less text and received most positive feedback. Most users did not read beyond the first line. Thus, providing users with more information to improve e-mail security seems to have no or even an adverse effect.

On the contrary, the "past correspondence" check was well received by all users. Some of them mentioned that they manually perform the same check in their own inbox, e.g. user O112: *"That will actually be useful, because I've found myself having multiple correspondence with the same person and I'll have to go and search for the name if I want to find something there. With that one, I think they're going to be much easier."* The provided functionality would save them time in real life, as they noted to regularly search for past e-mail correspondence with a given sender.

Other users indicated that the "past correspondence" check assured them of an e-mail's legitimacy, e.g. user O21: *"I just think it's fine, because there is a history of that e-mail, so it's fine."* This user noted that having exchanged e-mails before with a sender's e-mail address was a sign that the e-mail came from a trusted source. It did not necessarily matter to the user how many past e-mails were exchanged and about what. This mere fact indicated an established relation with the sender, which aligns with how users appraise their relation to a sender 4.1.2 and what has been described as "temporal embeddedness" as an indicator of trust [60]. Thus, users tended to ignore information that required more technical knowledge, but adopted information that augmented their existing e-mail processing behaviour.

4.2.2 Productivity versus security

In line with the previous theme, we found that users did not want to engage with features that interfered with their primary e-mail processing. This was most clearly observed with the "collegiate phishing report" nudge. Even if users did not read the provided content, we expected it to temporarily shift users' attention to the concept of e-mail legitimacy. In turn, this was expected to improve phishing detection. In most cases, however, users felt an urge to disregard or close the warning display as soon as possible when they saw it. For example, when user O16 in iteration 1 saw the warning nudge between the task ribbon and e-mails, they said *"'This e-mail was reported as suspicious by one of your colleagues.' Uh, did I do that? I use the cross."* They did not understand why they saw the nudge and closed it, without any further exploration.

In an attempt to increase user engagement with the nudge, iterations 2 and 3 showed the warning in a modal display in the newly loaded inbox. This was often experienced as highly disruptive. Yet, even in the last iteration when it was displayed as a highlighted e-mail that users had to click on themselves, most users only took a quick glance and closed it, e.g. user O14: *"So in this case, you have suspicion, my colleagues, with more information, uh, you missed a scheduled Zoom for blah, blah, blah. [. . .] Okay. Uh, I don't know. How can I get*

out of here?"

Some users who did read all the details in the "collegiate phishing report" nudge seemed to become more discerning throughout the task. Where they did not pay attention to the sender's e-mail address in previous inboxes, we found that they interpreted the sender e-mail address and explicitly mentioned whether a given e-mail was legitimate more often than before. It is unclear, however, for how long this effect would last. Together, these observations suggest that to improve users' secondary processing, we need to provide micro doses of information so not to harm users' primary processing.

4.2.3 User concerns on false positives

The "suspicion score" nudge aimed to alert users and enable better handling of suspicious e-mails. We found that users felt alerted and that most of them did not blindly delete or "junk" the given e-mail based on the warning. They often read the e-mail contents more carefully before deciding what action to take, which relates back to the content relevance theme in Section 4.1.1. While the majority of our users rated the "suspicion score" as the most useful feature, a few users expressed concerns around the possibility of a false positive warning. One of the phishing e-mails pretended to come from a Chinese company. When it contained the "suspicion score" nudge, user O211, of Chinese descent, mentioned that the warning was probably placed there due to algorithmic bias, as there is an allegedly large number of Chinese e-mail scams. They subsequently judged the e-mail as legitimate and ignored the recommended checks in the warning nudge that prompted users to double check the e-mail sender and links:

> *"**User O211:** This is Fujen International Education. This woman was. She is from China. [...] But I know the thing is, because I'm Chinese, I know lots of Chinese e-mails are flagged as not trustworthy, but this is just personal, so I'm going to forward it to my assistant instead of using this strong filter rating to decide [...] and we all know about algorithmic bias, do we?"*

With the increased attention for diversity and inclusion, users may grow especially sensitive to potential biases in automated decision systems. While the above e-mail was in fact a real phishing scam, it is important to take such user concerns into account when training detection systems and giving users security advice.

4.2.4 Ignorance toward security features

While the two nudges were ignored less, the "check" button remained untouched by 15 users until the researcher asked after three minutes if they saw the new button. Some users had seen it, but did not feel the need to explore it, e.g. user O114: *"[...] maybe I had seen it but I didn't really look at it and I didn't know what it was."* We found this surprising, as the button was located right next to the "Junk" button in the task ribbon in iterations 1–3 and next to the sender details in a different colour in iteration 4 to place it closer to the applicable e-mail content.

User O23 remarked *"Sometimes in the e-mails, you just go to, like, autopilot and you just... Yeah, I didn't even notice that."* This implies that new functionalities in e-mail UIs are easily missed, as users routinely process e-mails without thinking too much. Even after asking if users noticed the check button, not many consistently adopted its functionalities in the first three iterations. One explanation is that users did not find the information usable enough, as discussed in Section 4.2.1. This theme adds the possibility that users may have felt no need to use our features and preferred their own mitigation strategies when they suspected an e-mail (coded under "mitigation strategies"). These ranged from directly replying to the e-mail and judging by their response whether it was to be trusted, to asking colleagues and reporting it to IT. Together with Section 4.2.2, our findings suggest that users only adopt security features that align with their existing processing behaviour.

5 Discussion

As the sophistication of phishing attacks steadily grows [20], there is a pressing need to develop new e-mail security tools to protect users from falling for them. Here, we provide a formative evaluation of the usability of three e-mail user security tools based on the under-explored concepts of psychological nudges [27] and enhancing users' confidence in the legitimacy of genuine e-mails.

Through an iterative design process, we found that our past correspondence "check" button and "suspicion score" nudge help users detect phishing by affirming legitimate communication and alerting them of suspicious e-mails. These findings show the use of user-centric tool designs that enhance users' existing e-mail processing knowledge in a cost-effective way, instead of educating them about technicalities that many may find difficult to comprehend or easy to overlook. Future work could implement these designs into existing email clients and evaluate their effectiveness in-situ. Through these findings, we identified three usability versus security trade-offs, which we will discuss in light of developing usable e-mail security tools: highlighting legitimate vs. undesired communication (Section 5.1), supporting technical knowledge vs. existing behaviour (Section 5.2) and productivity vs. security (Section 5.3).

We also found that users largely process the same information found in e-mails (i.e., what the e-mail is about and whom the e-mail is from), but make judgement errors due to varying interpretations of those pieces of information—despite technical details explained in our tools. This is consistent with observations that users may notice surprising e-mail details,

but lack the knowledge to assess whether that information is suspicious [5, 35, 90]. Since all of our users are mandated to complete cybersecurity training (including phishing) every year, this was a surprising finding. It underlines the need for new approaches to improve detection [13, 38, 47, 57, 59].

5.1 Highlighting legitimate (desired) vs. undesired communication

The anti-phishing intervention literature typically focused on improving users' phishing detection ability by making them aware of suspicious cues in e-mails (see Section 2). Such "negative" framing is typically used in the wider usable security domain to alert users of potential security risks [37]. For most users, however, phishing e-mails likely comprise the minority of e-mails they receive. Our simulated inboxes therefore only contained 10–20% phishing. In these cases, users may rightfully assume that most e-mails are trustworthy, unless the prevalence of phishing e-mails is increased to a noticeable amount [66, 70]. This "prevalence paradox" limits the scope for user-centric anti-phishing tools within e-mail UIs, provided that user exposure to malicious e-mails remains relatively low. This implies that emphasising cues of legitimacy, such as with the past correspondence check, can maximise the scope to improve users' detection and confidence in e-mail legitimacy judgements.

Note that iterations 1–3 of the "check" tool explored multiple approaches to explain how to interpret technical sender and URL details: different buttons per functional content, simplifying language and information display, to no avail. As users were reluctant to process the provided details, we only kept the "past correspondence" check and then found positive user engagement with it. Although the "past correspondence" check could induce a false sense of security in cases of compromised or spoofed e-mail accounts, we believe the current findings provide a strong incentive to further explore e-mail user security tools that attend users to cues of legitimacy instead of phishing. For example, a next step could be to use language models to detect changes in the linguistic style of frequent senders to enhance the past correspondence check information and test it in an inbox that contains spoofed sender e-mails (of which we did not have examples).

In line with this paradigm and findings that many cybersecurity recommendations are unusable or vague [47, 57], our results further suggest that users would benefit from clearer organisational expectations on how to handle specific e-mail scenarios, whether legitimate or malicious. For example, tell users what to do with e-mails from free e-mail domains (e.g., Gmail). When users need to send or receive urgent requests, tell them what conventions to follow: e.g. to confirm with the person by phone or via a different channel than e-mail. This approach requires building a strong normative working culture that covers handling both legitimate and malicious e-mail communication. We would frame this contrast as "desired" versus "undesired" e-mail communication, instead of "phishing" versus "legitimate", as users may still have ill-defined ideas of what constitutes a phishing e-mail [20, 25].

5.2 Supporting (technical) knowledge vs. existing behaviour

Next, in support of Wash, Nthala, and Rader [83], we found that our most usable designs supported users' existing behaviour, whereas technical security-related content was ignored. Even within the "suspicion score" nudge, most users were sufficiently alerted by the mere presence of the nudge and did not read the recommended mitigation or suspicion score explanation. This aligns with recent quantitative findings from a large-scale study that more detailed warnings are not more effective than simple warnings [42], which highlights the power of our qualitative methods with a simulated environment. In the same vein, parsed URL and sender details shown with the original "check" button were also largely ignored. Users did not understand the utility of the provided information, even though the tool explained how to interpret the provided details. Users had a low tolerance for security information while processing e-mails.

These observations suggest that primary task interfaces may not be suitable for teaching users about e-mail security, which provides a further reason to refrain from using phishing simulations for "teachable moments" [42, 48, 65, 78]. Especially since all of our users are obliged to complete security and anti-phishing training and most of them still did not understand the provided security content, teaching users to accurately assess security information may be a task in vain. We therefore argue for an approach that leverages existing user behaviour, as was the case with the "past correspondence check" button and simple nudges like the "suspicion score".

Our finding that users were suspicious of e-mails when they perceived unexpected or "funny" content supports growing evidence that users pay most attention to e-mail content relevance, but not technical security indicators [32, 35, 52, 81, 83] such as URLs [90]. To then build on the idea of developing e-mail security tools that support existing user behaviour [83], another promising direction may be to categorise e-mails by their intent and show users nudges on e-mails with undesired discrepancies between intents and sender data. For example, when the e-mail message describes an "urgent request" and the sender is external to the user's organisation.

5.3 Balancing productivity vs. security

The last trade-off we found is how to balance users' productivity with security behaviour. The perception that security may be a (necessary) burden on users is a recurring theme [10, 14, 64] and we agree that security should not harm users' productivity. We also believe that well-designed security tools can

help users with minimal impact on their productivity, by leveraging how users actually reason about e-mails as described in Section 4.1, instead of explicitly trying to change users' (insecure) e-mail processing routines as suggested by [31]. The positive user engagement with the "past correspondence check" button exemplifies this. It relied on a heuristic that an e-mail can be trusted if there has been past correspondence with the sender's e-mail address. If there had not been any past correspondence, the button showed recommended actions to check the sender's legitimacy. This implicitly accords with our finding that users appraise their relation to the sender when viewing an e-mail. In contrast, the "collegiate phishing report" nudge was ineffective, as it was unclear to users how it related to the e-mails they wanted to process.

Another prominent observation is that most users never used the "Junk" button and several were confused about its functionality. This fits with findings that users' lack of understanding security concepts is associated with low or erroneous adoption of security tools [3, 63, 85]. Moreover, users did not distinguish between suspected phishing and generally unwanted e-mails, and just deleted both types. This possibly was the easiest action for them to understand and quickest to perform. We therefore expect that even with additional training, most users will refrain from using reporting functionalities, as these do not seem to align with how most users process e-mails and would affect their productivity. Even though the idea of personalised junk e-mail filtering systems has been suggested before [55, 89], security features that rely on machine learning models trained with users' processing behaviours (e.g., updating spam filters when users move e-mails to the junk folder [62]) may thus be unreliable. It may even be more beneficial to remove the "Junk" and "Report as phishing" buttons and "Junk" folder altogether. Taking into consideration recent studies [12, 51, 86], it is recommended that future research incorporates our suggested e-mail tool designs and assesses their efficacy in real-world settings. Furthermore, considering the themes and trade-offs highlighted in our research, there is ample opportunity to undertake more foundational UI design research pertaining to phishing.

5.4 Limitations

We did not test our features on live users or over longer time periods, as our formative evaluation aimed to first eliminate designs that provide low usability. In doing so, we tried to get as close to a realistic e-mail processing setting as possible by situating the study at an office desk in a regular office space, modelled the simulated inboxes after the institutional Outlook interface, participants participated at a time of their preference, and adapted e-mails received by colleagues at the participants' institute. Although the e-mails were unfamiliar to our participants, we did not expect this to fundamentally change how they reason about e-mails in the task. Indeed, our results without security features align with other qualitative works on how users process e-mails [32, 81, 83] and several participants remarked that the task felt like they were back at work. Lastly, we could not test if certain designs were better for certain demographic groups. Still, we would not expect significant differences by type of user, as we found a strong consensus among participants that the "suspicion score" nudge and "past correspondence" check were most useful.

6 Conclusion

Phishing is a persistent threat to organisations worldwide. Technical security measures alone do not sufficiently prevent people from falling for them, as users get exposed to new, evermore sophisticated attacks. Here, we provide a formative evaluation of three novel e-mail security tool concepts to help users discern trustworthy from suspicious e-mails in simulated inboxes. We used qualitative methods to gain a deep understanding of how our security tools affected user behaviour and thus why certain designs were (un)usable. Our "check" button supported user confidence in the trustworthiness of legitimate e-mails and the "suspicion score" nudge was deemed most useful. These findings highlight the potential of intuitive cues of legitimacy to augment existing user behaviour, instead of emphasising technical security knowledge. Together, they provide guiding principles for further usable security tool developments. We also found that users infer the trustworthiness of e-mails from the same types of information, but that differing interpretations of that information lead to erroneous judgements. This was surprising, as most of our users completed cybersecurity training before and have a technical study background. Future interventions that highlight desired versus undesired e-mail communication norms may help create more consensus among users. We hope these findings pave the way for a new generation of user-centric security tools to curb the risks of phishing threats.

7 Acknowledgements

We would like to thank the anonymous reviewers, Carlos Rombaldo Jr, Neil Amhis, Gerard Buckley and Nadine Michaelides for their valuable feedback on earlier versions of the manuscript. Sarah Zheng is supported by the UCL Dawes Centre for Future Crime, and Ingolf Becker is supported by the Engineering and Physical Sciences Research Council (grant number EP/W032368/1).

References

[1] Josh Aas et al. Let's encrypt: an automated certificate authority to encrypt the entire web. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, pages 2473–2487,

London, United Kingdom. Association for Computing Machinery, 2019. ISBN: 9781450367479. DOI: 10.1145/3319535.3363192.

[2] Jemal Abawajy. User preference of cyber security awareness delivery methods. *Behaviour & Information Technology*, 33(3):237–248, 2014. DOI: 10.1080/0144929X.2012.708787.

[3] Ruba Abu-Salma, M. Angela Sasse, Joseph Bonneau, Anastasia Danilova, Alena Naiakshina, and Matthew Smith. Obstacles to the adoption of secure communication tools. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 137–153, 2017. DOI: 10.1109/SP.2017.65.

[4] Devdatta Akhawe and Adrienne Porter Felt. Alice in warningland: a Large-Scale field study of browser security warning effectiveness. In *22nd USENIX Security Symposium (USENIX Security '13)*, pages 257–272, Washington, D.C. USENIX Association, 2013.

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

[6] Lukas Aldag, Peter Mayer, Mattia Mossano, Reyhan Duezguen, Bettina Lofthouse, Tatiana Landesberger, Melanie Volkamer, and Benjamin Berens. An investigation of phishing awareness and education over time: when and how to best remind users. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, 2020.

[7] Kholoud Althobaiti, Nicole Meng, and Kami Vaniea. I don't need an expert! Making URL phishing features human comprehensible. *Conference on Human Factors in Computing Systems - Proceedings*, 2021. DOI: 10.1145/3411764.3445574.

[8] Aurélien Baillon, Jeroen De Bruin, Aysil Emirmahmutoglu, Evelien Van De Veer, and Bram Van Dijk. Informing, simulating experience, or both: A field experiment on phishing risks. *PLoS ONE*, 14(12), 2019. ISSN: 19326203. DOI: 10.1371/journal.pone.0224216.

[9] Malak Baslyman and Sonia Chiasson. Smells phishy? an educational game about online phishing scams. In *2016 APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–11, 2016. DOI: 10.1109/ECRIME.2016.7487946.

[10] Adam Beautement, Ingolf Becker, Simon Parkin, Kat Krol, and M. Angela Sasse. Productive Security: A Scalable Methodology for Analysing Employee Security Behaviours. In Twelfth Symposium on Usable Privacy and Security (SOUPS 2016), Denver, CO. USENIX Association, 2016.

[11] Victoria Bellotti, Nicolas Ducheneaut, Mark Howard, Ian Smith, and R. Grinter. Quality versus quantity: e-mail-centric task management and its relation with overload. *Human-computer Interaction*, 20, 2005. DOI: 10.1207/s15327051hci2001&2_4.

[12] Frank Bentley, Josh Jacobson, Charlotte Sperling, Shiv Shankar, Chris Royer, and Ian McCarthy. Rethinking consumer email: the research process for yahoo mail 6. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, pages 1–6, Honolulu, HI, USA. Association for Computing Machinery, 2020. ISBN: 9781450368193. DOI: 10.1145/3334480.3375224.

[13] Benjamin Berens, Kate Dimitrova, Mattia Mossano, and Melanie Volkamer. Phishing awareness and education – when to best remind? In *Workshop on Usable Security and Privacy (USEC)*, 2022.

[14] Denis Besnard and Budi Arief. Computer security impaired by legitimate users. *Computers & Security*, 23(3):253–264, 2004. ISSN: 0167-4048. DOI: 10.1016/j.cose.2003.09.002.

[15] Marzieh Bitaab et al. Scam pandemic: how attackers exploit public fear through phishing. In *2020 APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–10, 2020. DOI: 10.1109/eCrime51433.2020.9493260.

[16] Jim Blythe, L. Camp, and Vaibhav Garg. In *Targeted Risk Communication for Computer Security*, pages 295–298, 2011. DOI: 10.1145/1943403.1943449.

[17] Virginia Braun and Victoria Clarke. One size fits all? what counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, 18(3):328–352, 2021. DOI: 10.1080/14780887.2020.1769238.

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

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

[20] Fiona Carroll, John Adejobi, and Reza Montasari. How good are we at detecting a phishing attack? investigating the evolving phishing attack email and why it continues to successfully deceive society. *SN Computer Science*, 3, 2022. DOI: 10.1007/s42979-022-01069-1.

[21] Marta E. Cecchinato, Abigail Sellen, Milad Shokouhi, and Gavin Smyth. Finding email in a multi-account, multi-device world. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 1200–1210, San Jose, California, USA. Association for Computing Machinery, 2016. ISBN: 9781450333627. DOI: 10.1145/2858036.2858473.

[22] Hongliang Chen, Christopher E. Beaudoin, and Traci Hong. Securing online privacy: An empirical test on Internet scam victimization, online privacy concerns, and privacy protection behaviors. *Computers in Human Behavior*, 70:291–302, 2017. ISSN: 07475632. DOI: 10.1016/j.chb.2017.01.003.

[23] Xi Chen, Indranil Bose, Alvin Chung Man Leung, and Chenhui Guo. Assessing the severity of phishing attacks: a hybrid data mining approach. *Decision Support Systems*, 50(4):662–672, 2011. DOI: 10.1016/j.dss.2010.08.020.

[24] Morton Deutsch. Trust and suspicion. *Journal of Conflict Resolution*, 2(4):265–279, 1958. DOI: 10.1177/002200275800200401.

[25] Julie Downs, Mandy Lanyon, and Lorrie Cranor. Decision strategies and susceptibility to phishing. In *Proceedings of the second symposium on Usable privacy and security (SOUPS)*, pages 79–90, 2006. DOI: 10.1145/1143120.1143131.

[26] Serge Egelman, Lorrie Cranor, and Jason Hong. You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In *SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2008. DOI: 10.1145/1357054.1357219.

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

[28] C. J. Gokul, Sankalp Pandit, Sukanya Vaddepalli, Harshal Tupsamudre, Vijayanand Banahatti, and Sachin Lodha. Phishy - A serious game to train enterprise users on phishing awareness. In *CHI PLAY 2018 - Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*, pages 169–181, 2018. DOI: 10.1145/3270316.3273042.

[29] Catherine Grevet, David Choi, Debra Kumar, and Eric Gilbert. Overload is overloaded: email in the age of gmail. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 793–802, Toronto, Ontario, Canada. Association for Computing Machinery, 2014. ISBN: 9781450324731. DOI: 10.1145/2556288.2557013.

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

[31] Jonas Hielscher, Annette Kluge, Uta Menges, and M. Angela Sasse. "taking out the trash": why security behavior change requires intentional forgetting. In *New Security Paradigms Workshop*, NSPW '21, pages 108–122, Virtual Event, USA. Association for Computing Machinery, 2021. ISBN: 9781450385732. DOI: 10.1145/3498891.3498902.

[32] Markus Jakobsson. The Human Factor in Phishing. *Privacy Security of Consumer Information*, 7:1–19, 2007.

[33] Jurjen Jansen and Rutger Leukfeldt. Coping with cybercrime victimization: an exploratory study into impact and change. *Journal of Qualitative Criminal Justice and Criminology*, 6(2):205–228, 2018.

[34] Jurjen Jansen and Paul van Schaik. The design and evaluation of a theory-based intervention to promote security behaviour against phishing. *International Journal of Human-Computer Studies*, 123:40–55, 2019. ISSN: 1071-5819. DOI: 10.1016/j.ijhcs.2018.10.004.

[35] Asangi Jayatilaka, Nalin Asanka Gamagedara Arachchilage, and Muhammad Ali Babar. Falling for phishing: an empirical investigation into people's email response behaviors, 2021. DOI: 10.48550/ARXIV.2108.04766.

[36] Matthew L. Jensen, Michael Dinger, Ryan T. Wright, and Jason Bennett Thatcher. Training to Mitigate Phishing Attacks Using Mindfulness Techniques. *Journal of Management Information Systems*, 34(2):597–626, 2017. ISSN: 1557928X. DOI: 10.1080/07421222.2017.1334499.

[37] Allen C Johnston and Merrill Warkentin. Fear appeals and information security behaviors: an empirical study. *MIS quarterly*:549–566, 2010. DOI: 10.2307/25750691.

[38] Iacovos Kirlappos and M. Angela Sasse. Security education against Phishing: A modest proposal for a Major Rethink. *IEEE Security and Privacy*, 10(2):24–32, 2012. DOI: 10.1109/MSP.2011.179.

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

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

[41] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Teaching johnny not to fall for phish. *ACM Transactions on Internet Technology*, 10(2), 2010. ISSN: 15335399. DOI: 10.1145/1754393.1754396.

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

[43] Eric Lin, Saul Greenberg, Eileah Trotter, David Ma, and John Aycock. Does domain highlighting help people identify phishing sites? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2075–2084, Vancouver, BC, Canada. Association for Computing Machinery, 2011. ISBN: 9781450302289. DOI: 10.1145/1978942.1979244.

[44] Xin (Robert) Luo, Wei Zhang, Stephen Burd, and Alessandro Seazzu. Investigating phishing victimization with the heuristic–systematic model: a theoretical framework and an exploration. *Computers & Security*, 38:28–38, 2013. ISSN: 0167-4048. DOI: 10.1016/j.cose.2012.12.003.

[45] Gloria Mark, Shamsi T. Iqbal, Mary Czerwinski, Paul Johns, Akane Sano, and Yuliya Lutchyn. Email duration, batching and self-interruption: patterns of email use on productivity and stress. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 1717–1728, San Jose, California, USA. Association for Computing Machinery, 2016. ISBN: 9781450333627. DOI: 10.1145/2858036.2858262.

[46] Marijn Martens, Ralf De Wolf, and Lieven De Marez. Investigating and comparing the predictors of the intention towards taking security measures against malware, scams and cybercrime in general. *Computers in Human Behavior*, 92(November 2018):139–150, 2019. ISSN: 07475632. DOI: 10.1016/j.chb.2018.11.002.

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

[48] Steven J. Murdoch and M. Angela Sasse. Should you phish your own employees? Bentham's Gaze. 2017. URL: https://www.benthamsgaze.org/?p=1756 (visited on 01/24/2023).

[49] James Nicholson, Lynne Coventry, and Pamela Briggs. Can we fight social engineering attacks by social means? assessing social salience as a means to improve phish detection. In *Proceedings of the 13th Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 285–298. USENIX Association, 2017.

[50] Adam Oest, Penghui Zhang, Brad Wardman, Eric Nunes, Jakub Burgis, Ali Zand, Kurt Thomas, Adam Doupé, and Gail Joon Ahn. Sunrise to sunset: Analyzing the end-to-end life cycle and effectiveness of phishing attacks at scale. *Proceedings of the 29th USENIX Security Symposium*:361–377, 2020. DOI: 10.5555/3489212.3489233.

[51] Soya Park, Amy X. Zhang, Luke S. Murray, and David R. Karger. Opportunities for automating email processing: a need-finding study. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, Glasgow, Scotland Uk. Association for Computing Machinery, 2019. ISBN: 9781450359702. DOI: 10.1145/3290605.3300604.

[52] Kathryn Parsons, Marcus Butavicius, Malcolm Pattinson, Agata McCormac, Dragana Calic, and Cate Jerram. Do users focus on the correct cues to differentiate between phishing and genuine emails? In *ACIS 2015 Proceedings - 26th Australasian Conference on Information Systems*, 2015. ISBN: 9780646953373.

[53] Kathryn Parsons, Agata McCormac, Malcolm Pattinson, Marcus Butavicius, and Cate Jerram. Phishing for the truth: A scenario-based experiment of users' behavioural response to emails. *IFIP Advances in Information and Communication Technology*, 405:366–378, 2013. ISSN: 1868422X. DOI: 10.1007/978-3-642-39218-4_27.

[54] Justin Petelka, Yixin Zou, and Florian Schaub. Put your warning where your link is: Improving and evaluating email phishing warnings. *Conference on Human Factors in Computing Systems*, 2019. DOI: 10.1145/3290605.3300748.

[55] Vipul Ved Prakash and Adam O'Donnell. Fighting spam with reputation systems: user-submitted spam fingerprints. *Queue*, 3(9):36–41, 2005. ISSN: 1542-7730. DOI: 10.1145/1105664.1105677.

[56] Emilee Rader and Anjali Munasinghe. "wait, do i know this person?": understanding misdirected email. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–13, Glasgow, Scotland Uk. Association for Computing Machin-

ery, 2019. ISBN: 9781450359702. DOI: 10.1145/3290605.3300520.

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

[58] Robert W. Reeder, Adrienne Porter Felt, Sunny Consolvo, Nathan Malkin, Christopher Thompson, and Serge Egelman. An experience sampling study of user reactions to browser warnings in the field. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, Montreal QC, Canada. Association for Computing Machinery, 2018. ISBN: 9781450356206. DOI: 10.1145/3173574.3174086.

[59] Benjamin Reinheimer, Lukas Aldag, Peter Mayer, Mattia Mossano, Reyhan Duezguen, Bettina Lofthouse, Tatiana von Landesberger, and Melanie Volkamer. An investigation of phishing awareness and education over time: When and how to best remind users. *Proceedings of the 16th Symposium on Usable Privacy and Security, SOUPS 2020*:259–284, 2020.

[60] Jens Riegelsberger, M. Angela Sasse, and John D. McCarthy. 53 Trust in Mediated Interactions. In *Oxford Handbook of Internet Psychology*. Oxford University Press, 2009. ISBN: 9780199561803. DOI: 10.1093/oxfordhb/9780199561803.013.0005.

[61] Fortune Jeff John Roberts. Exclusive: facebook and google were victims of $100m payment scam. 2017. URL: https://fortune.com/2017/04/27/facebook-google-rimasauskas/ (visited on 09/01/2022).

[62] Robert L Rounthwaite, Joshua T Goodman, David E Heckerman, John D Mehr, Nathan D Howell, Micah C Rupersburg, and Dean A Slawson. Feedback loop for spam prevention, 2007. US Patent 7,219,148.

[63] Scott Ruoti, Jeff Andersen, Scott Heidbrink, Mark O'Neill, Elham Vaziripour, Justin Wu, Daniel Zappala, and Kent Seamons. "we're on the same page": a usability study of secure email using pairs of novice users. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 4298–4308, San Jose, California, USA. Association for Computing Machinery, 2016. ISBN: 9781450333627. DOI: 10.1145/2858036.2858400.

[64] Angela Sasse, Sacha Brostoff, and D Weirich. Transforming the 'weakest link' — a human/computer interaction approach to usable and effective security. *BT Technology Journal*, 19, 2001. DOI: 10.1023/A:1011902718709.

[65] M. Angela Sasse and Steven J. Murdoch. Still treating users as the enemy: entrapment and the escalating nastiness of simulated phishing campaigns. Bentham's Gaze. 2021. URL: https://www.benthamsgaze.org/?p=3992 (visited on 01/24/2023).

[66] Ben D. Sawyer and Peter A. Hancock. Hacking the Human: The Prevalence Paradox in Cybersecurity. *Human Factors*, 60(5):597–609, 2018. ISSN: 15478181. DOI: 10.1177/0018720818780472.

[67] Stuart E. Schechter, Rachna Dhamija, Andy Ozment, and Ian Fischer. The emperor's new security indicators. In *2007 IEEE Symposium on Security and Privacy (SP '07)*, pages 51–65, 2007. DOI: 10.1109/SP.2007.35.

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

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

[70] Kuldeep Singh, Palvi Aggarwal, Prashanth Rajivan, and Cleotilde Gonzalez. Training to Detect Phishing Emails: Effects of the Frequency of Experienced Phishing Emails. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 63(1):453–457, 2019. ISSN: 2169-5067. DOI: 10.1177/1071181319631355.

[71] Cheryl B Stetler, Marcia W Legro, Carolyn M Wallace, Candice Bowman, Marylou Guihan, Hildi Hagedorn, Barbara Kimmel, Nancy D Sharp, and Jeffrey L Smith. The role of formative evaluation in implementation research and the queri experience. *Journal of general internal medicine*, 21(Suppl 2):S1, 2006.

[72] R.H. Thaler and C.R. Sunstein. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press, 2008. ISBN: 9780300146813.

[73] René van Bavel, Nuria Rodríguez-Priego, José Vila, and Pam Briggs. Using protection motivation theory in the design of nudges to improve online security behavior. *International Journal of Human Computer Studies*, 123:29–39, 2019. ISSN: 10959300. DOI: 10.1016/j.ijhcs.2018.11.003.

[74] Kami Vaniea, Lujo Bauer, Lorrie Faith Cranor, and Michael K. Reiter. Out of sight, out of mind: effects of displaying access-control information near the item it controls. In *2012 Tenth Annual International Conference on Privacy, Security and Trust*, pages 128–136, 2012. DOI: 10.1109/PST.2012.6297929.

[75] Kami Vaniea, Lujo Bauer, Lorrie Faith Cranor, and Michael K. Reiter. Studying access-control usability in the lab: lessons learned from four studies. In *Proceedings of the 2012 Workshop on Learning from Authoritative Security Experiment Results*, LASER '12, pages 31–40, Arlington, Virginia, USA. Association for Computing Machinery, 2012. ISBN: 9781450311953. DOI: 10.1145/2379616.2379621.

[76] Arun Vishwanath, Tejaswini Herath, Rui Chen, Jingguo Wang, and H. Raghav Rao. Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. *Decision Support Systems*, 51(3):576–586, 2011. ISSN: 01679236. DOI: 10.1016/j.dss.2011.03.002.

[77] Melanie Volkamer, Karen Renaud, Benjamin Reinheimer, and Alexandra Kunz. User experiences of TORPEDO: TOoltip-poweRed Phishing Email DetectiOn. *Computers and Security*, 71:100–113, 2017. ISSN: 01674048. DOI: 10.1016/j.cose.2017.02.004.

[78] Melanie Volkamer, Martina Angela Sasse, and Franziska Boehm. Analysing Simulated Phishing Campaigns for Staff. In *European Symposium on Research in Computer Security (ESORICS)*, pages 312–328. Springer, Cham, 2020. DOI: 10.1007/978-3-030-66504-3_19.

[79] Jaclyn Wainer, Laura Dabbish, and Robert Kraut. Should i open this email? inbox-level cues, curiosity and attention to email. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 3439–3448, Vancouver, BC, Canada. Association for Computing Machinery, 2011. ISBN: 9781450302289. DOI: 10.1145/1978942.1979456.

[80] Rick Wash. Folk models of home computer security. *ACM International Conference Proceeding Series*, 2010. DOI: 10.1145/1837110.1837125.

[81] Rick Wash. How Experts Detect Phishing Scam Emails. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW), 2020. ISSN: 25730142. DOI: 10.1145/3415231.

[82] Rick Wash and Molly M. Cooper. Who provides phishing training? facts, stories, and people like me. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, Montreal QC, Canada. Association for Computing Machinery, 2018. ISBN: 9781450356206. DOI: 10.1145/3173574.3174066.

[83] Rick Wash, Norbert Nthala, and Emilee Rader. Knowledge and capabilities that Non-Expert users bring to phishing detection. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 377–396. USENIX Association, 2021. ISBN: 978-1-939133-25-0.

[84] Zikai Alex Wen, Zhiqiu Lin, Rowena Chen, and Erik Andersen. What.hack: engaging anti-phishing training through a role-playing phishing simulation game. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, Glasgow, Scotland Uk. Association for Computing Machinery, 2019. ISBN: 9781450359702. DOI: 10.1145/3290605.3300338.

[85] Alma Whitten and J. D. Tygar. Why johnny can't encrypt: a usability evaluation of pgp 5.0. In *Proceedings of the 8th Conference on USENIX Security Symposium*, SSYM'99, page 14, Washington, D.C. USENIX Association, 1999.

[86] Oliver Wiese, Joscha Lausch, Jakob Bode, and Volker Roth. Beware the downgrading of secure electronic mail. In *Proceedings of the 8th Workshop on Socio-Technical Aspects in Security and Trust*, STAST '18, San Juan, Puerto Rico. Association for Computing Machinery, 2020. ISBN: 9781450372855. DOI: 10.1145/3361331.3361332.

[87] Min Wu, Robert Miller, and Greg Little. Web wallet: preventing phishing attacks by revealing user intentions. In *Proceedings of the second symposium on Usable privacy and security (SOUPS)*, volume 149, pages 102–113, 2006. DOI: 10.1145/1143120.1143133.

[88] Weining Yang, Aiping Xiong, Jing Chen, Robert W. Proctor, and Ninghui Li. Use of phishing training to improve security warning compliance: evidence from a field experiment. In *HoTSoS: Proceedings of the Hot Topics in Science of Security: Symposium and Bootcamp*, pages 52–61. Association for Computing Machinery, 2017. DOI: 10.1145/3055305.3055310.

[89] Seongwook Youn and Dennis McLeod. Spam decisions on gray e-mail using personalized ontologies. In *Proceedings of the 2009 ACM Symposium on Applied Computing*, SAC '09, pages 1262–1266, Honolulu, Hawaii. Association for Computing Machinery, 2009. ISBN: 9781605581668. DOI: 10.1145/1529282.1529565.

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

A Screenshots of design iterations

Figures 3, 4 and 5 below show the design updates for each iteration for each of the three feature concepts. The updates mainly regarded positioning within the inbox interface, while keeping the contents the same. Only the "check" button design was updated more significantly for the last iteration compared to the initial design.

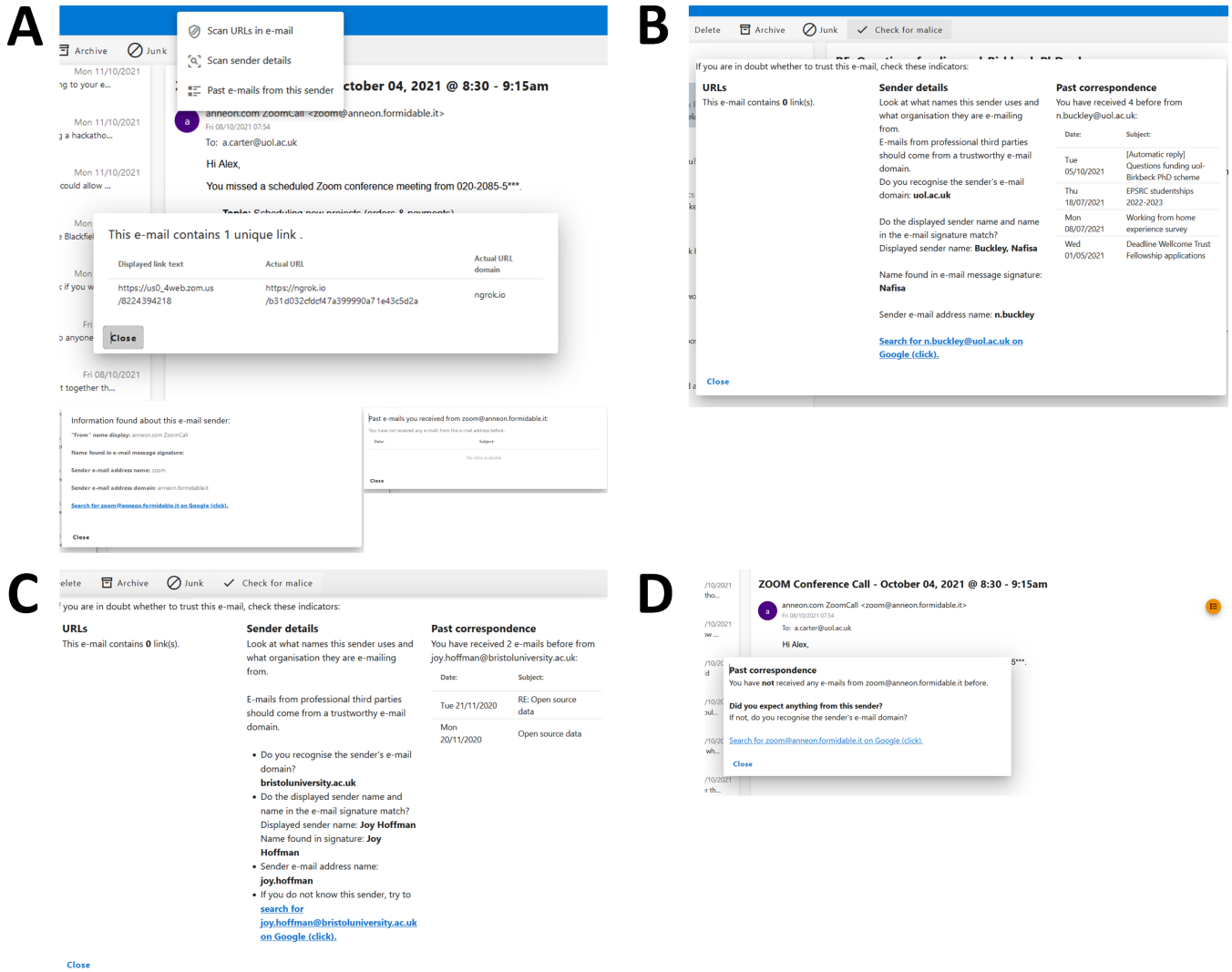


Figure 3: **"Check" button versions throughout the design iterations. A.** The first version of the button consisted of three sub menu items that appeared when a user hovered over the main button: (i) "Scan URLs in e-mail", which upon clicking would display an overlaid pop-up window with a table overview of all links found in the e-mail, the actual URL of the links and the actual URL domain; (ii) "Scan sender details", which displayed a list of sender name and e-mail address details as found in the user-facing e-mail header, as well as the e-mail body; (iii) "Past e-mails from this sender", which would display an overview of past e-mails received from the selected e-mail's sender e-mail address. **B.** After users pointed out the inefficiency of having three sub menu items to click in the first iteration's design, all three components from the first version were displayed at once when users clicked on the single "check" button. **C.** Users in the second iteration often found the displayed information too overwhelming (i.e., too much and/or too complicated) and tended not to read it. Hence, the amount of information was slightly reduced and displayed in a list for a more structured overview. **D.** In the last iteration, only the check for "past correspondence" was kept, since most users ignored or did not find the other technical information on URLs and sender details usable. We were aware of the potential false sense of security from this check in the case of spoofed e-mail addresses. Our task did not include spoofed sender phishing examples from the start, hence we used the last iteration to evaluate the usability of the concept of highlighting intuitive cues of legitimacy, rather than technical cues of malice. Also, to make the button more noticeable, it appeared as an orange icon next to the sender details.

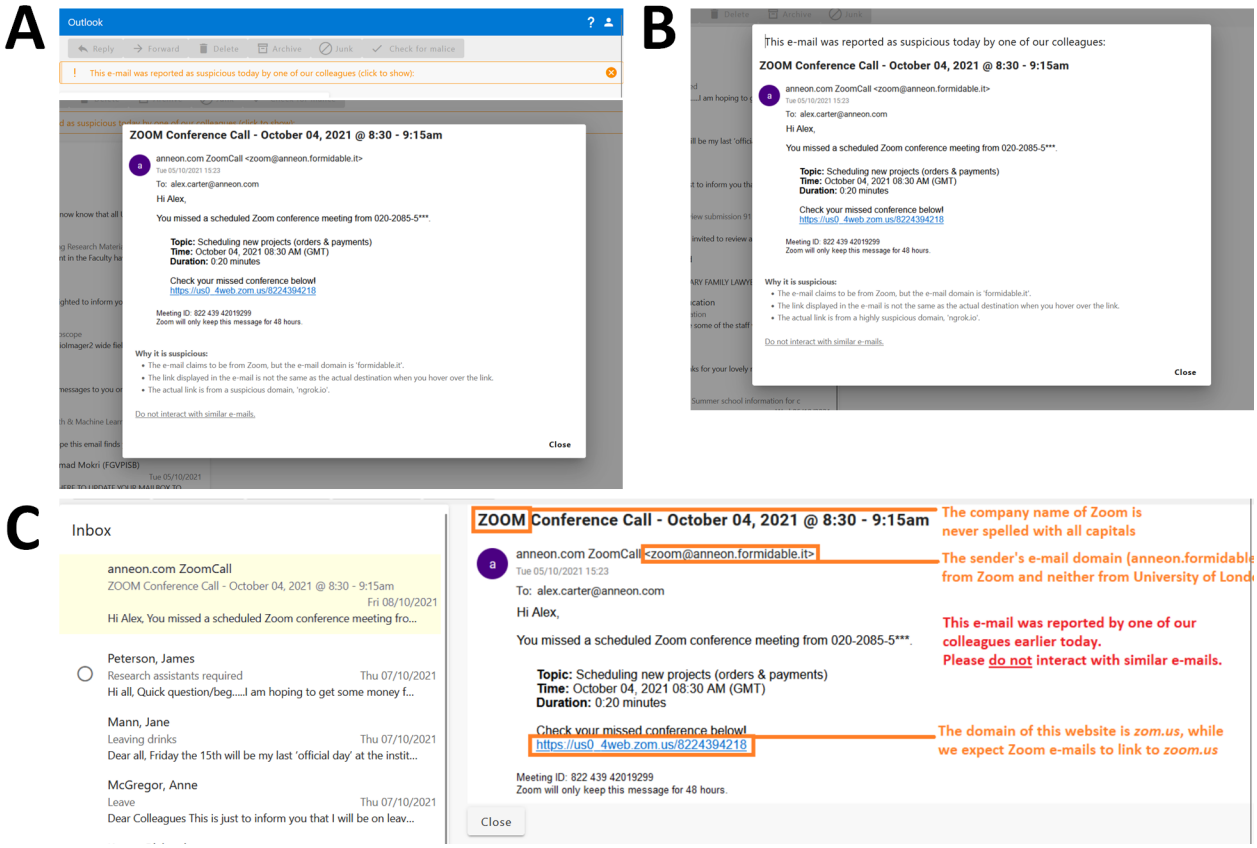


Figure 4: **"Collegiate phishing report" nudge versions throughout the design iterations A.** The first version was displayed as a warning banner between the task ribbon and the e-mails, which read "This e-mail was reported as suspicious today by one of our colleagues (click to show):" (left screenshot). If users clicked on it, a display on top of the inbox appeared with a phishing e-mail, a list with all reasons why it was malicious and that users should look out for similar e-mails (right screenshot). **B.** After many users either got confused, ignored or rapidly clicked away the warning banner in the first iteration, the updated version displayed the same display with the nudge text at the top when users loaded the new inbox. **C.** Feedback on the previous iteration indicated users' overall annoyance with the display. To make the nudge less disruptive, it was displayed as a highlighted e-mail at the top of the e-mails list. When users clicked on it, they saw the phishing e-mail with annotations in orange and the nudge text in red.

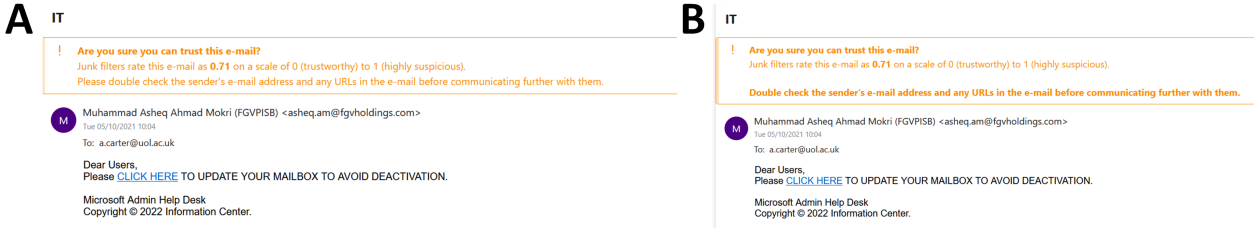


Figure 5: **"Suspicion score" versions throughout the design iterations. A.** Throughout the first three iterations, users were unanimously positive about the suspicion score design: an orange banner displayed on top of phishing e-mails with a score of 0.5 or higher. There were three lines of text. First a boldfaced line that warned users of whether they were sure they could trust the opened e-mail. Next, an explanation of why the warning is shown (high score on an automated suspicion scoring scale) and two recommended actions for the user (double checking links and sender details found in the e-mail). We showed the variability in suspicion scores for different suspicious e-mails to indirectly encourage users to think about the true legitimacy of a given e-mail. That is, to let them think of the possibility of misclassified "edge cases"—e-mails with relatively low suspicion socres (e.g. around .60). **B.** Users in previous iterations tended not to read the full warning text. Hence, the recommended actions were highlighted more by making them boldfaced and separating them with an extra line break for the last (fourth) iteration.

B Coding

Table 2 below shows all codes used to annotate users' reasoning and annotation frequency. The full code book including descriptions of each code is available via the OSF project page.

Top level	Secondary level	#
prioriti-sation approach	bottom-up	11
	quickly skim e-mails	42
	read whole e-mails	4
	senders-based	9
	top-down	40
	urgency	16
mitigation strategy	ask colleagues	3
	block sender	4
	blocked external sender content	1
	call sender	8
	check organisation via internet	13
	check past correspondence	2
	double check sender e-mail	14
	inspect linked page	3
	message actual internal sender	3
	not open attachment	1
	rely on antivirus to detect potential malice	2
	reply and evaluate response	8
	report to IT	1
	safe links	3
	train junk detection system	2
signals non-suspicious	formal e-mail signature	1
	internal e-mail	5
	IT Service Desk (ISD) checked	1
	looks important	3
	no warning sign	1
	not requesting sensitive information	2
	past correspondence	1
	proper written e-mail	3
	trusted sender e-mail address	15
	unclear reason	8
signals suspicious	external sender	26
	fear appeal	1
	funny URLs	21
	no online info about sender organisation	1
	non-professional sender e-mail address	41
	requesting personal details	10
	unclear reason	19
	unexpected or funny content	90
	unexpected sender or recipient name or e-mail	54
	urgent matter	3
	warning message	23
study feedback	design limitation	42
	unfamiliarity with context	14

Top level	Secondary level	#
processing reasons	assume known or trusted sender	46
	automated e-mail	8
	disseminate message	64
	from internal organisation	49
	high frequency	2
	important or urgent	37
	keep for reference	35
	meeting	113
	newsletter	4
	no action required	64
	no time for request	22
	not right audience or not personally targeted	36
	of personal interest	41
	outdated	12
	perform requested action or respond to query	84
	think about or research it before further action	41
	thread	21
	uninteresting or irrelevant	61
	unknown sender	8
intended process-ing actions	archive	66
	categorise in subfolder	14
	delete	157
	flag or pin	34
	forward	141
	junk	64
	leave in inbox	122
	reply	162
inter-vention feedback	functionality not clear	23
	missing useful info	11
	not applicable	1
	not sufficiently visible	2
	placement	3
	positive	15
	too much information to process	1
	unaware	17
	want to close nudge pop up asap	8
	warning could be false positive	4
prior experi-ences	experience academic context	1
	external sender warning	2
	known spam	8
	senders with unprofessional e-mail	1
	sensitised to security	2

Table 2: Final coding structure with reference frequency per code. Nine top-level codes and 86 secondary level codes were defined based on 27 session transcripts.