PhishED:
Automated contextual feedback for reported Phishing

Adam Jenkins
University of Edinburgh

Nadin Kökciyan
University of Edinburgh

Kami Vaniea
University of Edinburgh

Abstract

When people receive malicious emails that claim to be from a trusted entity like their bank, they can be frightened and uncertain about how safe it is to ignore or delete the message. The uncertainty is hard on users and it may lead them to engage in unsafe actions like clicking on links "just to check." In this work we look at how to provide quick and accurate support to people who report phishing so that they can confidentially take appropriate action. For this, we will build a phishing-advice tool, PhishEd, that allows people to report malicious emails that they encounter and get automatically generated advice in response that is contextual to the suspicious email. The advice is meant to both help them make an informed decision about the reported email as well as provide some education to help them in handling future malicious emails better.

1 Introduction

Phishing is one of the most effective vectors for tricking users into providing sensitive information, such as account details, to attackers who then use the information to harm users, organizations, and society at large. Recent attacks, such as the Colonial Pipeline shutdown in the US can be attributed to phishing [16]. In the UK, 86% of businesses reported receiving phishing attacks [8] and while 76% of UK users believe that they can recognize and avoid suspicious links in emails [5], our own research shows that only about 8% can accurately read a URL [1]. The staggering £53.7 million lost in the UK in 2020 to impersonation fraud also attests to the scale of the phishing problem faced by business and users [9].

Employer-provided training tends to be delivered either up-front [19] or given after the employee has fallen for a mock phishing attack [14]. These approaches do work, but they are very employer-centric in that employers choose the time and content of the training. Delivering the same training to everyone also forces a compromise between being comprehensive and containing detailed actionable guidance [11]. For example, training commonly advises users to "look at the destination of links" [18] but does not have the space to explain how to do so. Users in this situation have virtually no agency or control and lack the ability to shape their own learning to their needs. Users also experience phishing as an infrequent unexpected event that is mixed in with other activities. Phishing by its nature is designed to appear urgent or threatening. A user receiving such a message may be worried or scared about what might happen if they were to ignore it. Currently their only options are to delete the potential phish (scary) or report it and ask for advice, which may take some time or even not come at all [2].

In this project, we propose a novel phishing-advice tool, PhishEd, which accepts reports from users of potential phishing emails, uses artificial intelligence (AI) to parse out contextual phishing features, and quickly responds back to the user with advice that uses the content of the reported email to explain the reasoning or frame the decision the user needs to make. For example, a reported phishing might contain "HMRC" (HM Revenue and Customs - UK Government department for collection of taxes) but have links leading to Dropbox. The auto response would inform the user that the email is not from HMRC, using technical features such as DomainKeys Identified Mail (DKIM) signatures [7], and highlight that the links lead instead to Dropbox, which HMRC would never use. It would also provide examples from the email to evidence these claims. The objectives of this project are: (i) help users confidently make safe security decisions, (ii) harness "teachable moments" to demonstrate how to detect fraud and relevant technical skills, and (iii) encourage

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

end-users to want to continue report phishing in the future.

2 Background

Contextual Phishing Feature

In this project, our focus is using AI to provide phishing advice while making the reasoning process transparent to the user. In other words, PhishEd leverages AI's ability to extract and reason about contextual features of phishing in support of user decision making.

Example contextual phishing features that we identified in our previous work [2, 3] are as follows:

Contextual keywords A phishing email includes a set of keywords that may reveal the context of the message. For example, if a reported email uses terms like "shutdown", "email", and "account", the user may think that such a message was sent through an organization. Similarly, organization keywords like "PayPal", "HMRC", and "Dropbox" can indicate the implied organizational context of the email.

Uniform Resource Locators (URLs) A phishing email is designed to trick users to click on URLs included in the email. Such URLs can be analyzed in conjunction to other contextual factors to provide contextual guidance. For example, if an email contains the "PayPal" keyword but the URLs do not lead to PayPal's official website, that information can be provided to the user.

Email headers include meta-information about the sender. For example, the From address can be checked against DKIM signatures or whitelists of organization domains.

After examining the contextual features of a message, the phishing-advice tool will provide a content-based advice as a response to the user's reported email. For example: "This email is not from the organization's IT department which will only ever send email from the XXX@XXX.com email address.", while providing further justifications about the sender's identity, the content of the message and so on.

Embedded Training and Education

Education and training materials have long been considered as interventions for reducing the risks of phishing attacks, and hence have received interest from the academic community [12]. Training for end-users is available in range of formats and products. One of the most common and popular is that of embedded training, where simulated phishing attacks are conducted within the working contexts of user [14]. Although this form of training has been shown to be highly effective, implementing such training can take considerable effort to construct and monitor simulated phish and their results [2]. Additional interventions have been proposed such as the use of Serious Games [6, 21] or responsive solutions such as chat-bots [4].

Advice and training will routinely advise reporting suspicious emails [17], however research indicates that users may not do so due to a lack of confidence surrounding their abilities to spot legitimate phish [15]. The majority of phishing intervention research has focused on training users to identify phish, however this project intends to extend research by investigating the potential "teachable moment" once an suspicious email has been reported [10]. To our knowledge, the PhishEd project is the first investigation to focus solely on developing interventions for providing contextualized phishing advice on demand.

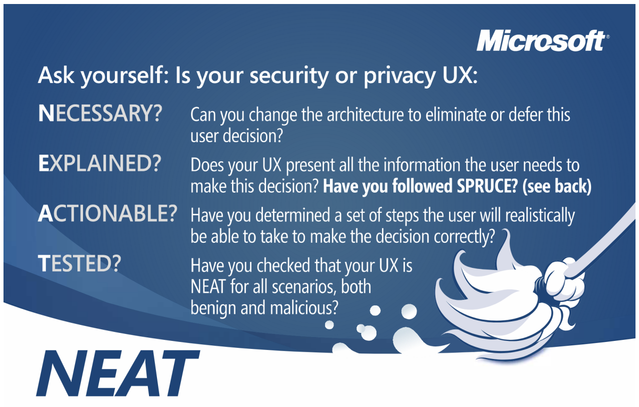
3 Design: Template & System

Nutrition Labels We have taken initial inspiration from the work of Kelly et al. who developed "privacy nutrition labels", essentially condensing privacy policies and their information into food nutrition label like format [13]. This approach was successfully adapted to the phishing domain by Althobaiti et al. [3] who developed phishing report interface to explain phishing features to humans, as depicted in Figure 3.

Masters Project A student Master thesis projects was conducted as part of this ongoing work, which focused on on creating initial designs for the PhishEd system. Zhang used a user-centered approach to inform their designs, implementing a prototype, gaining user feedback and further iterating to create a final mock design [22] shown in Figure 4.

NEAT & SPRUCE We plan to improve on our template designs by integrating Microsoft's security warning design principles, NEAT (Figure 1) and SPRUCE (Figure 2). This guidance was developed to be a part of Microsoft's product teams' development process, even providing wallet-sized cards and lessons on integrating the advice in practice [20]. We use this guidance to provide us with an initial design space for generating contextual advice templates for discussion in our planned focus groups (Section 4).

Artificial Intelligence (AI) AI has historically focused on the automatic detection of phishing emails using machine learning approaches, such as classification, that promise high accuracy. In the phishing context, it is not trivial to reach 100% accuracy due to the changing nature of the problem (e.g., attackers change their strategies constantly) and the high costs of false positives. Our focus is using AI to provide phishing advice while making the reasoning process transparent to the user. PhishEd leverages AI's ability to extract and reason about contextual features of phishing in support of user



Microsoft

Ask yourself: Is your security or privacy UX:

NECESSARY? Can you change the architecture to eliminate or defer this user decision?

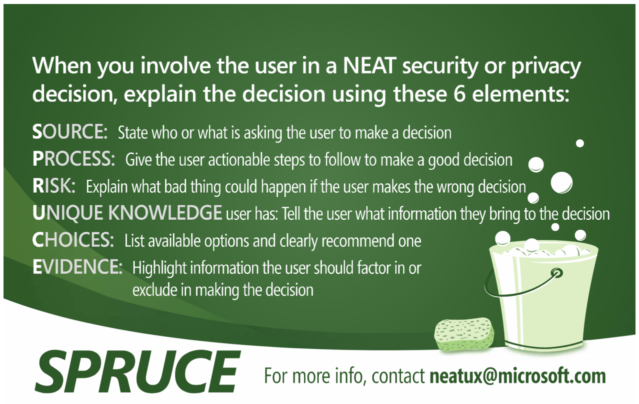
EXPLAINED? Does your UX present all the information the user needs to make this decision? **Have you followed SPRUCE? (see back)**

ACTIONABLE? Have you determined a set of steps the user will realistically be able to take to make the decision correctly?

TESTED? Have you checked that your UX is NEAT for all scenarios, both benign and malicious?

NEAT

Figure 1: Microsoft's NEAT



When you involve the user in a NEAT security or privacy decision, explain the decision using these 6 elements:

SOURCE: State who or what is asking the user to make a decision

PROCESS: Give the user actionable steps to follow to make a good decision

RISK: Explain what bad thing could happen if the user makes the wrong decision

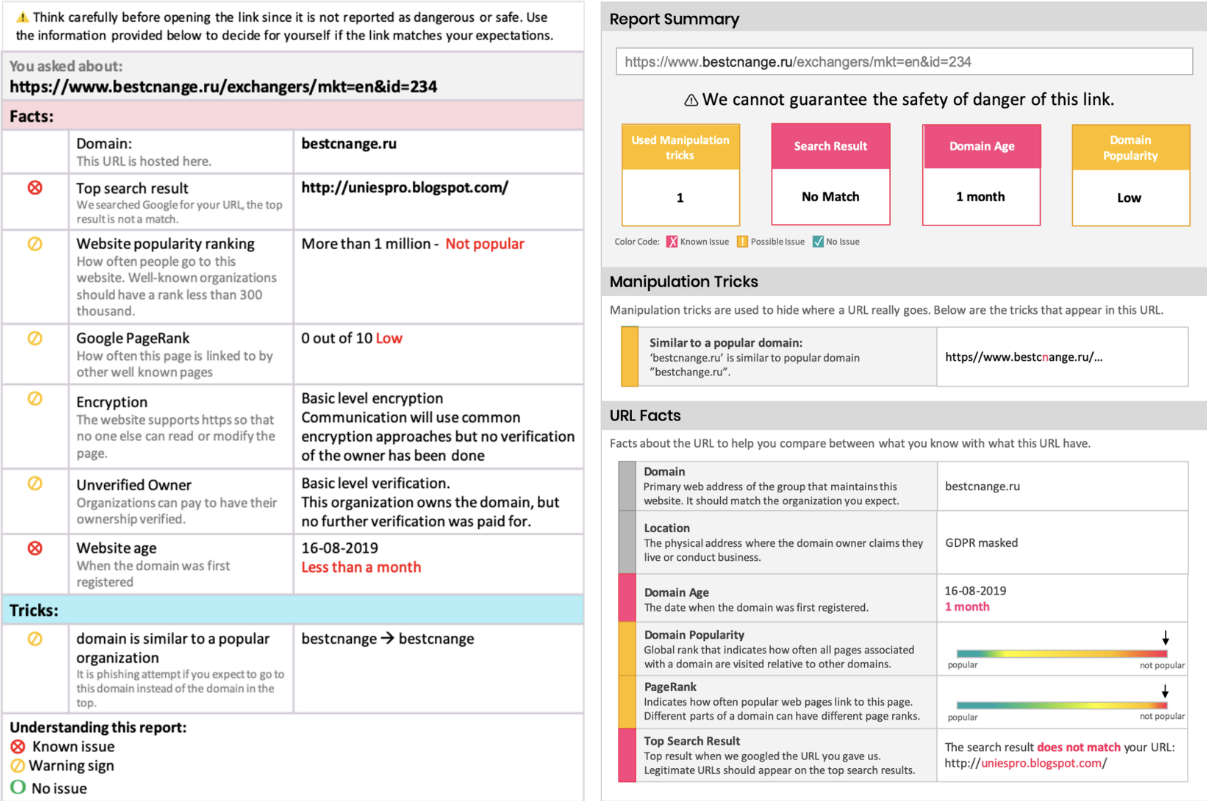
UNIQUE KNOWLEDGE user has: Tell the user what information they bring to the decision

CHOICES: List available options and clearly recommend one

EVIDENCE: Highlight information the user should factor in or exclude in making the decision

SPRUCE For more info, contact **neatux@microsoft.com**

Figure 2: Microsoft's SPRUCE



⚠ Think carefully before opening the link since it is not reported as dangerous or safe. Use the information provided below to decide for yourself if the link matches your expectations.

You asked about:
https://www.bestcnange.ru/exchangers/mkt=en&id=234

Facts:

	Domain: This URL is hosted here.	**bestcnange.ru**
⊗	Top search result We searched Google for your URL, the top result is not a match.	**http://uniespro.blogspot.com/**
⊘	Website popularity ranking How often people go to this website. Well-known organizations should have a rank less than 300 thousand.	More than 1 million - Not popular
⊘	Google PageRank How often this page is linked to by other well known pages	0 out of 10 Low
⊘	Encryption The website supports https so that no one else can read or modify the page.	Basic level encryption Communication will use common encryption approaches but no verification of the owner has been done
⊘	Unverified Owner Organizations can pay to have their ownership verified.	Basic level verification. This organization owns the domain, but no further verification was paid for.
⊗	Website age When the domain was first registered	16-08-2019 Less than a month

Tricks:

⊘	domain is similar to a popular organization It is phishing attempt if you expect to go to this domain instead of the domain in the top.	bestcnange → bestcnange

Understanding this report:
⊗ Known issue
⊘ Warning sign
○ No issue

Report Summary

https://www.bestcnange.ru/exchangers/mkt=en&id=234

⚠ We cannot guarantee the safety of danger of this link.

Used Manipulation tricks	Search Result	Domain Age	Domain Popularity
1	No Match	1 month	Low

Color Code: ☒ Known Issue ! Possible Issue ☑ No Issue

Manipulation Tricks

Manipulation tricks are used to hide where a URL really goes. Below are the tricks that appear in this URL.

Similar to a popular domain: 'bestcnange.ru' is similar to popular domain "bestchange.ru".	https//www.bestcnange.ru/...

URL Facts

Facts about the URL to help you compare between what you know with what this URL have.

Domain Primary web address of the group that maintains this website. It should match the organization you expect.	bestcnange.ru
Location The physical address where the domain owner claims they live or conduct business.	GDPR masked
Domain Age The date when the domain was first registered.	16-08-2019 **1 month**
Domain Popularity Global rank that indicates how often all pages associated with a domain are visited relative to other domains.	popular — not popular
PageRank Indicates how often popular web pages link to this page. Different parts of a domain can have different page ranks.	popular — not popular
Top Search Result Top result when we googled the URL you gave us. Legitimate URLs should appear on the top search results.	The search result **does not match** your URL: http://uniespro.blogspot.com/

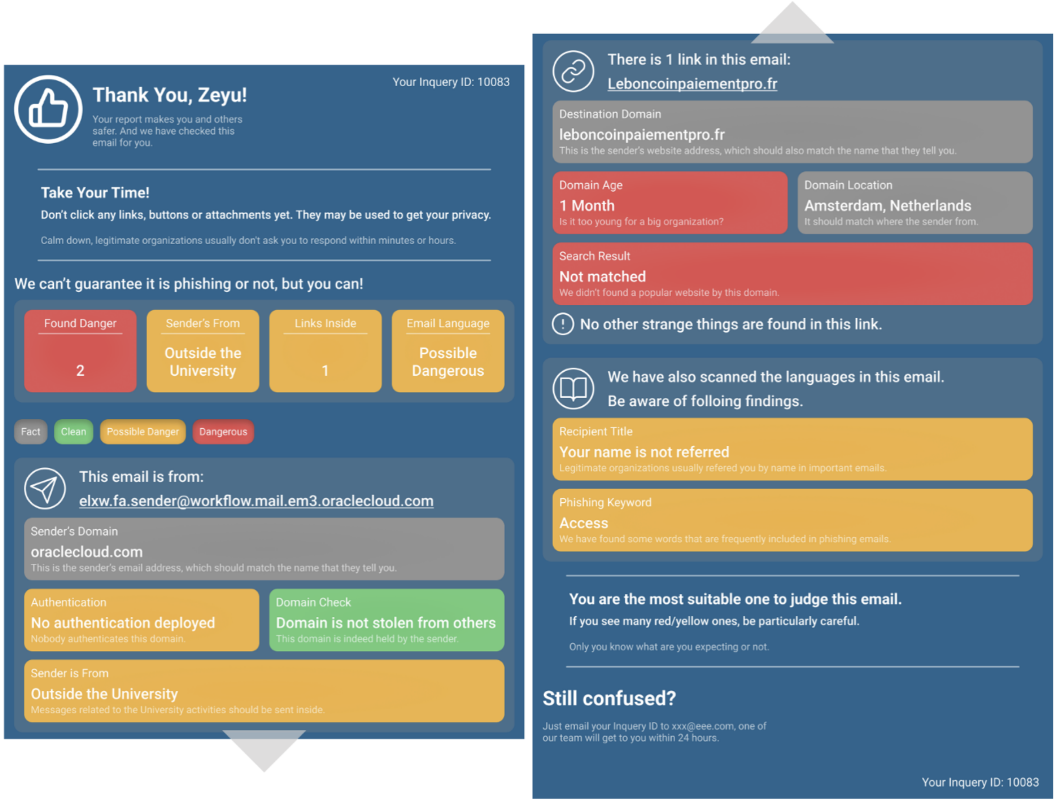
Figure 3: Reports designed by Althobaiti et al. [3] to breakdown features of phishing URLs.

decision making. We will use Natural Language Processing techniques such as Named-Entity Recognition or Semantic Role Labelling to capture contextual features; and unsupervised machine learning approaches such as clustering to group emails together based on extracted features. AI approaches will aid in identifying ongoing phishing campaigns, and our campaign-based advice templates will be used to provide feedback to the user when they report potential phish.

4 Future Work: Iteration and Deployment

To achieve the goals of PhishEd, we will employ a range of methods to aid in the iterative design process, and the final evaluation of the designs.

Focus Groups & Design Workshops Once further ideation and design alternatives have been completed we will take



Your Inquery ID: 10083

Thank You, Zeyu!

Your report makes you and others safer. And we have checked this email for you.

Take Your Time!

Don't click any links, buttons or attachments yet. They may be used to get your privacy.

Calm down, legitimate organizations usually don't ask you to respond within minutes or hours.

We can't guarantee it is phishing or not, but you can!

Found Danger	Sender's From	Links Inside	Email Language
2	Outside the University	1	Possible Dangerous

Fact · Clean · Possible Danger · Dangerous

This email is from:
elxw.fa.sender@workflow.mail.em3.oraclecloud.com

Sender's Domain
oraclecloud.com
This is the sender's email address, which should match the name that they tell you.

Authentication
No authentication deployed
Nobody authenticates this domain.

Domain Check
Domain is not stolen from others
This domain is indeed held by the sender.

Sender is From
Outside the University
Messages related to the University activities should be sent inside.

There is 1 link in this email:
Leboncoinpaiementpro.fr

Destination Domain
leboncoinpaiementpro.fr
This is the sender's website address, which should also match the name that they tell you.

Domain Age
1 Month
Is it too young for a big organization?

Domain Location
Amsterdam, Netherlands
It should match where the sender from.

Search Result
Not matched
We didn't found a popular website by this domain.

No other strange things are found in this link.

We have also scanned the languages in this email.
Be aware of folloing findings.

Recipient Title
Your name is not referred
Legitimate organizations usually refered you by name in important emails.

Phishing Keyword
Access
We have found some words that are frequently included in phishing emails.

You are the most suitable one to judge this email.

If you see many red/yellow ones, be particularly careful.

Only you know what are you expecting or not.

Still confused?

Just email your Inquery ID to xxx@eee.com, one of our team will get to you within 24 hours.

Your Inquery ID: 10083

Figure 4: Initial design of automated responses by Zeyu Zhang (Masters Thesis in Design Informatics) [22].

these initial designs and templates to end-users for user-driven iterative design. The project aims to identify the types and structure of phishing advice that users desire when reporting suspicious emails, by using the contextual information contained within the suspected phish. These studies will be conducted with members of our institution, including both staff and students from a range of departments and backgrounds to generate a representative participant user group.

Lab Study Upon completing our initial designs we will conduct a small lab-based user study using a mock email inbox and a set of tasks designed to encourage engagement with our reporting system. The lab study is focused on investigating the legitimacy of our early designs and prototypes, while also being used to identify and usability concerns and the suitability of our proposed solution within the working contexts of users.

Longitudinal Study We are currently working closely with an organization to incorporate our proposed system into their phishing reporting systems. We will develop an Outlook add-in to integrate PhishEd to Microsoft Outlook, where users will be able to report phish easily. This add-in will be deployed for the users who participate in our study. We will monitor the usage of the reporting tool and also measure the tool's impact on the participants' experiences when reporting and handling suspicious emails.

Acknowledgments

This work was funded in part by the UKRI Strategic Priorities Fund via the REPHRAIN Research Centre, and a Google Research Award. The authors would also like to thank Zeyu Zhang for his contributions to the project and appreciate the continual feedback of the TULiPS research group.

References

[1] Sara S. Albakry, Kami Vaniea, and Maria K. Wolters. What is this url's destination? empirical evaluation of users' url reading. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–12, April 2020.

[2] Kholoud Althobaiti, Adam Jenkins, and Kami Vaniea. A case study of phishing incident response in an edu-

cational organization. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 2021.

[3] Kholoud Althobaiti, Nicole Meng, and Kami Vaniea. I don't need an expert! making url phishing features human comprehensible. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, May 2021.

[4] Kholoud Althobaiti, Kami Vaniea, and Serena Zheng. Faheem: Explaining URLs to people using a slack bot. In *2018 Symposium on Digital Behaviour Intervention for Cyber Security (AISB)*, pages 1–8, Liverpool, UK, April 2018. University of Liverpool.

[5] Lloyds Bank. *UK Consumer Digital Index 2019*. 2019.

[6] Gamze Canova, Melanie Volkamer, Clemens Bergmann, and Benjamin Reinheimer. Nophish app evaluation: lab and retention study. In *NDSS workshop on usable security*, 2015.

[7] Dave Crocker, Tony Hansen, and Murray Kucherawy. Domainkeys identified mail (dkim) signatures. *ser. RFC6376*, 2011.

[8] Media Department for Digital, Culture and (UK GOV) Sport. *Cyber Security Breaches Survey 2020: Statistical Release*. 2020.

[9] UK Finance. *Fraud - The Facts 2021*. 2022.

[10] 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, August 2021.

[11] Cormac Herley. So long, and no thanks for the externalities: the rational rejection of security advice by users. In *Proceedings of the 2009 workshop on New security paradigms workshop*, pages 133–144, 2009.

[12] Jason Hong. The state of phishing attacks. *Commun. ACM*, 55(1):74–81, jan 2012.

[13] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W Reeder. A" nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, pages 1–12, 2009.

[14] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Teaching johnny not to fall for phish. *ACM Trans. Internet Technol.*, 10(2), jun 2010.

[15] Youngsun Kwak, Seyoung Lee, Amanda Damiano, and Arun Vishwanath. Why do users not report spear phishing emails? *Telematics Informatics*, 48:101343, 2020.

[16] Phil Muncaster. Colonial pipeline incident sparks 'help desk' phishing attacks. *InfoSecurity Magazine*, 2021.

[17] Phishing attacks: dealing with suspicious emails and messages. http://bit.ly/3tTwQpC, December 2018. Accessed Feb. 2022.

[18] Phishing.org. 10 ways to avoid phishing scams, 2022.

[19] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. How I learned to be secure: a census-representative survey of security advice sources and behavior. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, October 24-28, 2016*, pages 666–677, Vienna, Austria, 2016. ACM.

[20] Rob Reeder, E Cram Kowalczyk, and Adam Shostack. Helping engineers design neat security warnings. In *Proceedings of the Symposium On Usable Privacy and Security (SOUPS), Pittsburgh, PA*, 2011.

[21] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason I. Hong, and Elizabeth Nunge. Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. In *Proceedings of the 3rd Symposium on USAble Privacy and Security, SOUPS*, volume 229, pages 88–99, Pittsburgh, Pennsylvania, USA, July 2007. ACM.

[22] Zeyu Zhang. *Designing an Autoresponder for Phishing Email Reports*. PhD thesis, University of Edinburgh, 2021.