Let the weakest link fail, but gracefully

Let the weakest link fail, but gracefully: understanding tailored phishing and measures against it

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van
de rector magnificus prof. dr. S.K. Lenaerts,
voor een commissie aangewezen door het College
voor Promoties, in het openbaar te verdedigen op
woensdag 24 januari 2024 om 13:30 uur

door

Pavlo Burda

geboren te Horodok, Oekraïne.

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

Voorzitter:	prof. dr. E.R. van den Heuvel
Promotoren:	prof. dr. S. Etalle
	dr. N. Zannone
Copromotor:	dr. L. Allodi
Leden:	prof. F. Paci (University of Verona)
	prof. G. Russello (Universtiy of Auckland)
	prof. dr. A. Serebrenik

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

The work in this dissertation has been partially funded by Rijksdienst voor Ondernemend Nederland (RVO) under the the ITEA3 programme through the DEFRAUDIfy project (grant no. ITEA191010)

The work in this dissertation has been partially funded by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) under the the INTERSCT project (Grant No. NWA.1162.18.301) and the SeReNity project (Grant No. cs.010)

Printed by: ADC Nederland - 's-Hertogenbosch

Cover design: Valentyn Kudlai

A catalogue record is available from the Eindhoven University of Technology Library.
ISBN 978-90-386-5934-3

An electronic version of this dissertation is available at `http://repository.tue.nl` and `http://www.pavlo.it`

Summary

Humans play a critical role in computer systems, making them an integral part of their attack surface. Social engineering attacks specifically aim to deceive individuals to gain unauthorized access to sensitive information or deploy malware on their systems. The most common form of social engineering attack is phishing, by which an attacker sends fraudulent messages (typically in the form of emails), claiming to be from a reputable and trusted source. Phishing and, more in general, social engineering attacks exploit inherent vulnerabilities rooted in human cognition, allowing attackers to manipulate system users in executing actions against their own self-interest. Since these vulnerabilities are universal among potential targets and cannot be easily fixed (e.g., by training), they present a consistent and relatively stable attack surface for attackers to exploit. This allows attackers to minimize the complexity and costs associated with deploying malware-based attacks, while still potentially achieving a high impact on the system. Phishing attacks are evolving rapidly and increasing in sophistication: attackers can gather targeted information about their victims and use it to build tailored phishing attacks to further improve attack efficacy. The gathered information, such as contextual information on the targets and their environment, can be used to craft believable pretexts that significantly increase the attack success rates. The variability of attack characteristics (pretext, links) and resemblance to regular communication make most detection attempts and user anti-phishing education largely ineffective. The potential scalability and relatively low effort to deploy a tailored phishing campaign create significant risks for Internet users, organizations, and institutions; historical examples include financial losses, data breaches, and disruption of democratic processes.

Because of the multidisciplinary nature of social engineering, there is a lack of a structured and coherent understanding of the complex socio-technical mechanisms that underpin it. As generic, mass phishing is considered the most prevalent form of social engineering attacks, empirical research has so far mainly focused on these 'untargeted' phishing attack scenarios. However, the nuances involved in targeted phishing attacks and the effects of the manipulation of information relevant to the target remain unexplored. Further, existing countermeasures lag behind the evolution of more sophisticated phishing attacks, such as tailored phishing. We thus examine the following main research question:

What are the current gaps in our understanding of tailored phishing attacks from the target, attacker, and defender perspectives, and which technological and organizational methods can be employed to address these gaps?

To answer this question, we develop a framework to structure and map social engineering attacks to a high-level representation of relevant human cognitive processes. The framework, grounded on existing well-established cognitive theories, is used to carry out a systematic literature review of the extant empirical research, allowing us to identify gaps in relation to

experiment characteristics, core cognitive features, and the exploitable attack surface from the target perspective.

We then adopt the attacker's perspective and investigate what techniques can be best exploited in a tailored attack, and their effects on human cognition with a field experiment in two large organizations. This provides insights into the relationship between cognitive exploits, their delivery methods, and the organizational settings. Current countermeasures, such as automated detection and training, might be off-target for such sophisticated attacks. As such, we investigate the defender perspective by exploring technological and organizational mitigation strategies. We develop a novel approach, as a browser extension, to support users in detecting phishing websites by identifying which website a phishing web page is imitating using a mix of automated textual and visual features recognition techniques. The second mitigation approach targets organizational environments whereby user reporting of attacks to the IT department of an organization may be a significant, yet untapped, resource to mitigate advanced campaigns. We employ qualitative and quantitative methods to investigate what influences reporting behavior 1) by interviewing employees targeted in a simulated tailored phishing attack at a small IT company, and 2) by investigating the intention to report as a function of certain human factors. Our findings shed light on the rationale and motivation of users reporting phishing attacks and provide a more comprehensive understanding of traits and attitudes affecting individuals' cyber security behaviors. This carries a series of implications on both theoretical and practical levels that can help organizations to improve their security processes, anti-phishing training, and awareness programs.

In the context where the functioning of our society heavily depends on digital communications, this thesis advances social engineering research by identifying, estimating and mitigating the associated risks. We identify open gaps in research by contextualizing social engineering attacks in the cognitive sciences domain. We estimate the potential risks by demonstrating how target-related information in phishing can overrun the effects of conventional phishing. Finally, we mitigate the risks by showing why humans – the targets of such attacks – can be the current best defense against, otherwise unstoppable, sophisticated phishing attacks.

We do not spontaneously learn that we don't learn that we don't learn.

Nassim Nicholas Taleb

Contents

List of Figures

List of Tables

Acronyms

AMT Amazon Mechanical Turk

APT Advanced Persistent Threat

CE Central Executive

CERT Computer Emergency Response Team

CSS Cascading Style Sheets

DOM Document Object Model

HCI Human-Computer Interaction

HTML HyperText Markup Language

IP Internet Protocol

ISP Infromation Securiy Policy

LTM Long-Term Memory

MFA Multi-Factor Authentication

NGO Non-Governmental Organization

NIST National Institute of Standards and Technology

OCB Organizational Citizenship Behaviors

OSINT Open Source Intelligence

QR-code Quick-Response code

SE Social Engineering

SME Small-Medium Enterprise

SNS Social Networking Sites

SOC Security Operation Center

SSL Secure Sockets Layer

UI User Interface

URL Uniform Resource Locator

WM Working Memory

YoS Years of Service

1

Introduction

The term 'Social Engineering' (SE) is used in information security to refer to a type of attack wherein an attacker manipulates individuals to compromise the confidentiality, integrity, and availability of data and processes by exploiting human vulnerabilities [368]. These attacks act on and exploit the cognitive domain of individuals and societies, i.e., the set of human information processing mechanisms, to deceive targets and attain an impact on the physical world [238]. By deceiving humans behind the systems, attackers avoid complex and costly system-based exploitation, such as, scanning networks and engineering malware or exploits [15]. Real world consequences of such attacks may be the acquisition of assets and disruption of processes in the information and communication domains for financial, political, ideological and other motives.

1.1. Scope

SE attack types

Attackers can exploit a variety of communication channels to deliver their attacks, such as email, voice calls, instant messaging or even physical displacement of USB drives with malicious software. Phishing has become the most prevalent attack type, with the email as the main attack vector due to its wide adoption for professional and private communication; other common attack types include SMShing which can occur via SMS, vishing over voice calls etc. [66, 303]. SE attacks can be thought of as mechanisms by which the user is deceived into facilitating a security breach, such as revealing a password or downloading malware. For example, typical techniques can deliver a legitimate-looking link or attachment in an email, spoof the sender address, abuse the software User Interface (UI), or craft believable pretexts with the aim of persuading the target to disclose sensitive information or execute malicious code [147]. Furthermore, these mechanisms can vary in sophistication and scope where message, website and/or any file attachment is difficult to distinguish from legitimate sources and appear to be relevant for the targets' environment [59]. It is thus clear that there can be a multitude of possible attack types that can vary in terms of employed media (e.g., voice calls [339], QR codes [354]), attack artefacts (e.g., URLs [88], email attachments [380]),

content type (e.g., pretexts [377], contextualisation [256]), targetization levels (e.g., generic mass campaigns vs. spear-like phishing [61, 150]), and other variables that depend on the attack goals, attacker efforts and targeted users.

Human cognition

Regardless of the attack type, SE techniques exploit human characteristics that determine our emotional and mental constitution, our behavior and our social coexistence. Examples include curiosity, fear and trust, as well as social aspects as loyalty or respect for authority. From the attacker point of view, these characteristics can be skillfully leveraged, akin to software vulnerabilities in computer systems, to manipulate system users. Cognitive sciences provide a large body of evidence for disparate cognitive mechanisms that affect how humans process information through thought, experience and perception [27, 146]. Crucially in the context for this thesis, some of the cognitive processes ingrained in the human cognitive system manifest byproducts that, when triggered, can lead to undesirable decisional outcomes. For example, heuristics are fast and unconscious psychological rules that aid judgment and decision making, such as breaking under a red-light while driving or to perform repetitive tasks, that lead to appropriate decisions under most circumstances [146, 176, 340]. In certain cases, however, heuristics can lead to systematic errors or cognitive biases [74, 124, 175]. An example of heuristics' effect is scarcity, where the deceiver makes calls for urgent action to trigger the individuals' tendency to inflate the importance of an object or event when it is available for a limited time (such as seasonal sales) [74, 233]. These *cognitive vulnerabilities* are often triggered and exploited in SE attacks to deceive the target in performing an action against their intentions or best interest [256, 359, 375, 379]. Notably, Cialdini's definition [74] of cognitive triggers have been widely adopted in multiple domains, such as marketing [76], behavior change [296], governance [98] and in SE research as 'persuasion techniques' or 'principles' [9, 114]. Such tactics and combinations thereof are exploited by attackers across various attack vectors, such as emails, to bypass technical barriers and compromise the security of systems. This created the (unjust) saying of users as 'the weakest link' in cyber security.

As a result, research in SE is related to a variety of disciplines such as sociology, psychology and human computer interaction (HCI) among others, and deals with the interplay between technical aspects of an attack and the cognitive dimensions characterising its human element. To navigate this complex socio-technical problem, a new strain of empirical research investigating the interplay between *cognitive effects* and *attack features* emerged in the scientific literature [52, 238, 276, 320, 321]. For example, extant research investigated the influence of persuasion techniques inside email content on attack success [379], or how human attention can be manipulated to influence clicks on various system pop-ups [242]. However, there is no easy way to capture multiple perspectives from a variety of disciplines. Indeed, the multidisciplinary nature of this problem poses challenges in pinpointing research gaps, unanswered questions, and interpreting findings in empirical SE research.

Tailored phishing

Among various types of SE attacks, phishing stands out in terms of prevalence, impact and efficacy. Due to this relevance in the overall threat landscape, phishing has been extensively studied in the scientific literature [303]. Phishing usually entails un-targeted mass campaigns of one-time email messages aimed at large chunks of internet users exploiting a consistent arsenal of cognitive vulnerabilities. Because the overwhelming majority of phishing attacks in the wild are of this type, many empirical results in the extant literature are related to un-targeted, un-sophisticated attacks scenarios. On the other hand, in recent years there have been reports of increased volume of targeted phishing attacks [39, 40, 108, 352], commonly called spear-phishing, in which the target and their context are investigated so that the email is tailored to the receiver. Personal details or other relevant information is used to craft legitimate-looking messages and increase the likelihood of success. It is believed that spear-phishing emails are successful because personalisation creates trust and this contextual approach may deceive people who would otherwise be cautious about phishing attempts [54]. For instance, an attack on military personnel may involve an invitation to a retirement party for a general, prompting the recipients to click on a confirmation link [152]. More sophisticated attacks can be diluted over multiple iterations and reconnaissance steps to leverage the gathered information and avoid detection. For example, studies report multi-stage attacks against representatives of minorities in China where the language and topic of emails were highly tailored to the targets [47], or attacks against white-collar workers on LinkedIn where attackers, behind a fictitious company, exploited applicants' profiles to adapt their communication to the intended victims [15]. Although the effectiveness of spear-like attacks is broadly considered greater than their generic counterpart [150], the associated costs can be higher than with the latter; hence, such attacks are more commonly employed by resourceful actors, such as Advanced Persistent Threats (APTs) and state-sponsored groups, or when the return on investment is high enough [66, 108].

This creates a 'spectrum' of phishing attacks, whereby attacks fall between the two extremes of 'generic' phishing (un-targeted, large numbers and 'hit-or-miss' fashion) and spear phishing (targeting individuals or small groups, prior reconnaissance, multiple attacks steps), and vary in adaptation (e.g., to a large demographic, a specific organization or individual), in goals (e.g., harvesting credentials, delivering malicious files) and levels of sophistication (e.g, supporting infrastructure, detection avoidance) [61]. As any class of phishing, from generic to spear-phishing, still attempts to match the interest of the target to some degree, it remains challenging to characterize the 'tailoring degree' of an attack. Among the possible variants between generic and spear phishing, tailored phishing is becoming increasingly relevant to the overall threat landscape [15]. Similarly to generic phishing, tailored phishing is a one-time message attack targeting a relatively wide target group; however, the message is tailored to the target in a spear-like fashion. Tailored phishing can involve a basic 'hit-or-miss' strategy (as in un-targeted phishing), but employing more advanced open source intelligence (OSINT) and personalization techniques to leverage information about the targeted victims and/or their organization (as in spear-phishing). Such instances of phishing can be deployed at the scale of large organizations [326], but carrying the potential to be as effective as the more targeted 'spear' variants, in part, due to the growing capabilities of automated OSINT tools [110] and the availability of specialized toolkits to deliver large scale spear phishing

training campaigns [281]. Crucially, tailored phishing shares spear-phishing techniques, but at the scale of organizations or large groups of individuals, and this can lead to high impact with relatively low effort from the attacker.

Extant research on targeted variants of phishing primarily addresses user susceptibility [120, 142, 223, 320] and training effectiveness against spear-phishing [62, 66, 121, 194]; some investigate effects of target and email characteristics [209]. However, this research is still inconclusive on how variables, such as attack techniques, adaptation levels or user context, can affect the efficacy of tailored attacks in various situations [132, 320]. Indeed, there is still little understanding of the nuances of exploitation of target-relevant information and the associated effects on cognition which, in turn, hinders the development of effective countermeasures.

Countermeasures

Phishing research and development mainly focused on understanding, preventing and detecting incoming attacks. Following the NIST *protect*, *detect* and *respond* framework [252], state of the art phishing countermeasures integrate 1) training and awareness campaigns to 'immunize' users against the attacks [10], 2) signature matching and anomaly detection systems [11, 320] and 3) security operation centers and incident response procedures at organizations to mitigate possible impact [188]. Preventing users from falling into phishing attacks can be achieved by training users to identify the specific features of these attacks and increasing their awareness of the threat. Previous research has shown that training and awareness campaigns can have a significant impact in reducing the susceptibility of individuals to (generic) phishing attacks [167]. However, small portions of subjects may still be vulnerable and the same results may not be seen with spear-phishing attacks: some studies showed either no significant differences between generic and spear-phishing training effects [194], no effect of training at all [66] or marginally significant effects [62]. Anyhow, other studies have shown that awareness campaigns may often fail to appropriately change user behavior due to their inadequate implementation [35] and that the effectiveness of training inevitably decreases over time [31, 52]. As prevention methods cannot eliminate the risk of users falling for (spear) phishing attacks, software detection methods are employed to block them from reaching the user in the first place. The most common method for detecting phishing attacks is artifact filtering, which checks email content and URLs [180], and is similar to anti-spam filters. Different methods, such as data mining, machine learning, heuristics, and allow/block-lists, have been used to implement automated phishing detection [116]. Each solution has its own trade-offs with respect to accuracy, coverage and complexity among others. Notably, machine learning approaches have become popular, but they can still lead to a relatively high number of false positives (with high numbers of emails or websites to check, even a small error rate can get in the way of usability), require constant retraining to identify unknown attacks (such as spear-phishing), and are not universally applicable to protect all classes of targets (i.e., such solutions are often narrowed towards a predetermined target list of organizations or brands) [77, 148, 152, 347, 348]. To address the drawbacks of automated detection, studies have investigated crowdsourcing and reputation systems to improve reliability [117, 212, 241], but these solutions still face limitations against previously unseen attack instances [152, 348]. When both prevention and detection

are insufficient, response strategies can be put in place to lower the impact of an attack. The response can occur immediately after detection, for example, when users report an attack, or after the attack has already caused damage [188, 350]. The detection and response to spear-phishing attacks can be handled by security analysts in an organization's IT department or in a security operation center (SOC) where the remediation process may involve intercepting the attack and blocking traffic or investigating rogue domains. However, the short duration of these attacks can hinder this process [61, 161, 192] and containment procedures are sensitive to delays [152, 351], especially with unknown attacks [188]. Despite these efforts being able to thwart a consistent fraction of 'generic' phishing campaigns, no definitive solution to phishing, let alone tailored or spear phishing, has been found yet. Recent empirical research in SE is driving a revived push for mixed detection-response approaches that account for human-centric defense strategies, for example, by leveraging human intelligence to assist automation for response procedures [196] or by supporting individuals to avoid attacks with combined detection approaches [58, 362].

In this thesis, we adopt the perspective of human cognition to investigate the interaction of social and technical components of the SE phenomenon. Our efforts are aimed at exploring and understanding the interplay between the cognitive effects and attack features in SE attacks. In particular, we will focus on the challenges of empirical SE research when studying tailored phishing attacks. Further, we explore, investigate and propose possible mitigation strategies from a human-centric perspective in the attempt to challenge the widespread notion in computer security of 'humans as the weakest link'.

1.2. Research Gap

Lack of structured and coherent understanding of SE attacks

The research community in SE is lacking a structured, shared understanding of the dimensionality of the SE problem. This makes it particularly difficult to identify gaps and open research questions as well as to interpret experimental results. We have seen that the success of a SE attack depends on an often unpredictable combination of socio-technical factors. For example, attacks may consider the impersonation of a trusted entity, such as a government agency, company or affiliated contacts [200]. This can be achieved, for instance, by leveraging the weaknesses of a communication medium to spoof the sender of a message, such as an email address or a phone number [147]. Here, a technical factor is at play to enable the attack and raise the likelihood of success. At the same time, the attack success depends on the pretext – the purpose alleged in the message – which can leverage social, psychological and contextual factors, such as social norms, human emotions or timing, to make the target perform the desired action [148, 365, 375]. This interplay of factors is a major obstacle when studying and simulating SE attacks. On top of that, extant research in computer science has mainly focused on technical factors to investigate SE and countermeasures whereby, for example, technical solutions are developed on the communication protocol specifications, such as DKIM and SPF[1]. Social and psychological factors, on the other hand, have been

[1]DKIM (DomainKeys Identified Mail) and SPF (Sender Policy Framework) are two email authentication methods used to prevent email spoofing and verify the authenticity of email messages.

considered less often in computer security, although they relate to a large array of effects on human cognition and behavior which are no less important than the technological ones. The study of these factors commonly pertains to fields that include, but not limited to, information systems, social sciences and cognitive sciences [238]. In between, we can find disciplines such as HCI that somewhat bridge multiple fields. However, the overall efforts to comprehensively study the socio-technical aspects related to the SE domain remain so far relatively unstructured. It is therefore clear that the lack of a structured and coherent understanding of SE attacks poses important challenges for SE research, hindering the development of effective strategies to counter these attacks in practice.

Little understanding of effects of target-relevant information in phishing attacks

Since the majority of empirical research focuses on 'un-targeted' attack scenarios, there is a limited comprehension of the nuances involved in targeted attacks and the manipulation of information relevant to the target. Research on targeted attacks is carried out mainly on what is commonly defined as spear-phishing: this class of phishing broadly includes any type of phishing that explicitly targets individuals or small groups and that is somewhat personalized [147, 200]. However, this research is mostly focused on the recipient, with few experiments on how the attack artefacts (e.g., the message) and the situation (e.g., the user context) influence susceptibility [54, 320]. We have seen that target-related information can be manipulated in many ways to shape user behavior, such as matching user context and message pretext [132] or adding more or less details to a phishing email [62]. It is therefore still unclear what techniques can be best exploited in *tailored* attacks in various situations.

Similarly, there is little to no understanding of the associated effects on cognition in tailored attacks. Previous work in this direction mainly addressed the usage of persuasion techniques (i.e., the triggers of cognitive vulnerabilities) and the influence of human factors on user susceptibility (e.g., personality, knowledge or experience), mostly, in the context of generic phishing. For example, some persuasion techniques appear to be more popular than others [9, 114] and their relative efficacy varies [115, 379]. Short and long-term human factors can influence user susceptibility at various degrees and directions, for instance, high cognitive effort lowers phishing susceptibility [361], as opposed to curiosity which can increase the likelihood to fall for phishing [239]. The only studies on cognitive effects of *tailored* attacks report that certain persuasion techniques were more effective within spear-phishing messages [63], that their relative efficacy varies as a function of individual's life interests [209] and that attention, rather than cognitive effort, influences individual's response to spear-phishing emails [364]. Overall, there are disparate cognitive mechanisms at play that condition how attack artefacts and context affect an individual's processing and ultimately the susceptibility to a phishing attack [238]. This suggests that the efficacy of well-known cognitive effects in tailored attacks is unclear and the scientific evidence is still inconsistent.

Common phishing countermeasures are unsuited against tailored attacks

State of the art countermeasures integrate awareness campaigns and training of employees, advanced detection software and security operation centers. However, existing countermeasures are lagging behind the evolution of sophisticated phishing attacks, such as tailored

phishing [13]. Attack characteristics such as pretext and links are extremely variable, hindering the majority of detection attempts or generating too many false positives [148]. Further, the resemblance of these attacks to regular communication make training and awareness campaigns largely ineffective to 'immunize' a significant fraction of the victim pool [62]; anomalies in the communication still exist, such as unusual references to internal processes in an organization, but these are hard to formalize and cannot be captured automatically by a single technological solution. Therefore, organizations often rely on response teams, such as SOCs and CERTs[2], as the last line of defense [188]. Still, current containment procedures based on after-the-fact analyses are too slow to match the high velocity at which targeted phishing campaigns are known to affect their targets [161, 188, 351]. This brings us to the point where, albeit the significant progress in countering mass phishing campaigns, we are currently facing uncharted territory with limited knowledge when tackling tailored, spear-like phishing attacks.

1.3. Research Questions

The overreaching objective of this thesis is to shrink the gap in understanding social engineering attacks, specifically in terms of tailored phishing and countermeasures against it. Given that the success of SE attacks is affected by attack features, target characteristics and the countermeasures in place, we need to account for all three perspectives in our efforts. We thus articulate our work around the following main research question:

Main RQ: *What are the current gaps in our understanding of tailored phishing from the target, attacker, and defender perspectives, and which technological and organizational methods can be employed to address these gaps?*

The multidisciplinarity of the SE domain makes it particularly difficult to identify gaps and open research questions, and to interpret experimental results. This is particularly evident when dealing with more complex forms of SE, such as tailored phishing attacks. To answer the Main RQ, therefore, we must first draw a picture of the current state of empirical SE research.

We have seen that efforts to comprehensively study the socio-technical aspects related to the SE domain are so far relatively unstructured. For instance, much of phishing research focuses on the recipient, with less efforts to understand how target-related information, message pretext or user context influence SE susceptibility. Overlooking the effects of such factors may lead to conflicting findings, such as less effective training recommendations and imprecise risk assessments. The same can be said about the human cognitive factors influencing target deception, for example, with the measurement of how individuals process messages with persuasion techniques or how a contextual situation around the target affects user attention. As a result, it becomes difficult to reconcile the different findings and reach a comprehensive understanding of the 'SE attack surface'. For that, we need an instrument that can characterize the SE attack surface to evaluate and contextualize research results, and identify said gaps. To tackle this challenge, we ask the following question:

[2]SOC - Security operation Center and CERT - Computer Emergency Response Team

RQI: *How can we characterize the SE attack surface to evaluate and contextualize research results and identify gaps in empirical SE research?*

To answer this question, we study the interplay between the human cognitive processes and attack features in SE attacks. We develop and showcase a cognitive framework for characterizing SE attacks based on theories and models of human cognition drawn from the field of cognitive sciences. Thus, we should be able to better understand the relationships between features of SE attacks and results in empirical SE research. This means that we could compare different (simulated and non) SE attacks based on their cognitive features and reason over why or how was an attack (in-)effective in prompting a target to compliance. For example, we could identify shortcomings of experiments and simulated attacks, such as isolating factors that are difficult to recognize without a reference to the features of human cognition (e.g., effects of pretext and context) or reasoning over the attack adaptation to the targets (e.g., matching of attacker assumptions and target-related information). This is valuable to identify open problems in empirical SE research in relation to the aspects of human cognition. Hence, we ask the question:

RQII: *What are the open gaps between the features of human cognitive processes and empirical research in SE, including future research directions?*

To answer this question, we carry out a systematic literature review in the field of empirical SE research, focusing on experimental characteristics and core cognitive features from both attacker and target perspectives. One of the key findings is that the exploitable SE attack surface appears much larger than the coverage provided by the current body of research: for example, the effects of different pretexts and varied targetization levels are overall marginally investigated, and populations at risk of tailored attacks, such as in industry and institutional domains, are under-represented in research.

We also observe that one of the most relevant attack vectors in empirical SE research remains the email where un-targeted phishing constitutes the overwhelming majority of experiments. On top of that, phishing attacks are evolving rapidly and increase in sophistication: attackers can gather targeted information about their victims, and use it to build tailored attacks to further increase their efficacy. The gathered information, such as contextual information on the targets and their environment, can be used to craft believable pretexts, forge identities or even tune the tone of the decoy message. Whereas the effect of tailoring on attack success can be significant, which techniques can be best exploited in a tailored attack, and their effects on human cognition remain largely unexplored. As tailored phishing keeps growing into a significant threat, the question of what strategies can help mitigate it remains unanswered as well. Therefore, we ask the question:

RQIII: *How effective are tailored phishing campaigns in deceiving targets to perform an action, and what strategies can be employed to mitigate these attacks?*

To answer this question we perform a simulated tailored phishing campaign against two organizations, a mid-sized Dutch university and a large international company. In the experiment, we evaluate the effectiveness of tailored phishing across the two target domains and measure the velocity at which targets fall for the attack. Our findings reveal insights on the relation between tailored phishing, cognitive attacks, their delivery methods, and the

organizational settings of the targets.

To counter phishing attacks that target a specific user base, technical (e.g., detection software) and organizational (e.g., prevention and response) approaches can be employed. As in our tailored phishing campaign, phishing attacks commonly deliver URL payloads linking phishing pages that impersonate a target service, such as a company web-mail, to decoy users into revealing their online credentials. For our experiment, the velocity at which users are compromised suggests that detection should occur as early as possible. As a potential technical countermeasure, we develop and evaluate a method to identify the intended target of a 'zero-hour' phishing page (i.e., a new and unknown attack instance) by relying on its visual similarity with the original website, with the goal to support users in their decision-making. The key feature of this approach is that no predefined reference list of targets to protect is necessary. This can be valuable against attacks impersonating and targeting less popular brands or organizations which might not have the resources to defend against more advanced attack, such as tailored phishing. Our efforts provide a step forward towards the mitigation of phishing attacks, including targeted variants, as well as enable further experimentation in this direction.

We also observe, however, that current defensive strategies, including detection techniques and user training, may be off-target for sophisticated attacks. Further analysis of our tailored phishing campaign leads to the finding that user reporting of attacks to the IT department of an organization may be a significant, yet untapped, resource to enable the mitigation of advanced campaigns. Response teams at IT departments or operational centers can indeed uncover advanced campaigns more effectively than common detection techniques. Still, the efficiency of the reporting process depends on the number and the quality of user notifications. We argue that among the employees of an organization, there are some that are particularly good at detecting phishing. However, only a few users typically report phishing emails, and the rationale and incentives behind this are still unexplored in the scientific literature. To answer the second part of RQIII, we interview the university employees that reported our tailored phishing attack to the IT department and reconstruct the mental models reflecting the decision process of the participants. The preliminary results suggest that exploring users' reactions and decision making can shed light on how to improve and leverage phishing reporting to mitigate the impact of sophisticated attacks, such as tailored or spear-phishing. Therefore, we ask the question:

RQIV: *What rationale do users follow when deciding to report a phishing attack and what influences their decisions?*

To address this question, we investigate the user reactions to a simulated tailored phishing attack at a small IT enterprise in the Netherlands. By interviewing the targeted employees, we observe the engagement of a strong community reaction around the attack whereby several employees sent a warning to their colleagues triggering a 'collective defense' mechanism. Employees' answers reveal their reasoning behind detecting and (not) reporting the phishing email (e.g., the tailored nature of the attack which provoked curiosity in certain employees to investigate further), along with their thoughts and emotions.

Finally, to understand what other factors influence phishing reporting, we carried out a online survey to evaluate the relationships of certain human factors and intention to report phishing. We develop and test a research model of how personality traits, beliefs and atti-

tudes towards the organization and colleagues affect positive cyber security behaviors and, specifically, intention to report. The results of both interviews and survey reveal a series of implications for research and practice that extend our comprehension of possible organizational mitigation strategies against tailored phishing attacks.

1.4. Thesis Outline and Contributions

The objective of this work is to provide an answer to the Main RQ. The Main RQ is divided in four research questions that address its various aspects. We structure our results in this dissertation over three parts:

- Part I answers RQI and RQII that characterize human SE attack surface and open problems in SE in relation to human cognition
- Part II answers RQIII which concerns the effectiveness of tailored phishing attacks and potential strategies to mitigate these attacks
- Part III answers RQIV regarding the phishing reporting mechanisms as strategy to counter sophisticated attacks

Each part contains chapters that are based on one or multiple research publications that have been organized following a common research line. The original version of each publication has been slightly adapted to the thesis format. As each chapter is based on an independent article that needs to be self-contained, there may be some redundancy across the manuscript. Every chapter is introduced and the relation with the other ones is briefly described.

We begin with Chapter 2 where we provide an explanation of the main concepts useful to better understand the following chapters. Here we introduce background notions on SE attack phases and characteristics, the cognitive processes at play during target exploitation and various methods to counter SE attacks. Parts of this chapter appear in refereed conference publications.

In Chapter 3 we answer RQI by exploring the cognitive processes involved in a SE attack and developing a cognitive framework to characterize SE attacks based on theories and models of human cognition drawn from the field of cognitive sciences. The framework is meant to characterize and study SE attacks of varying types (e.g., via email, social networks) and complexity (e.g., increasing tailoring towards the targets) thus helping to contextualize and evaluate research results in SE (e.g., by providing a structure to isolate cognitive effects). To showcase the framework's application, we analyze two academic experiments simulating SE attacks and two real SE attacks of increasing sophistication to illustrate how the framework can be used to identify gaps and ways forward. This chapter is published in the proceedings of a peer-reviewed conference [57].

In Chapter 4, to answer RQII, we identify and characterize open gaps in empirical SE research via a systematic literature review. Our criteria cover the experiment setup, the characteristics of the simulated SE attack, the target's cognitive processes and characteristics, and the interactions between such variables. Our study shows that most experiments only partially reflect the complexity of real SE attacks and that the exploitable SE attack surface appears

much larger than the coverage provided by the current body of research. For example, despite their high relevance for both attack design and defense, factors such as targets' context and cognitive processes are often ignored or not explicitly considered in experimental designs. Similarly, the effects of different pretexts and varied targetization levels are overall marginally considered. We find that the literature is largely focused only on a few experimental setups, it lacks a common reference for attack targetization and the experimental outcomes are relatively inconsistent in defining when a SE attack is deemed successful. Finally, we report promising, interdisciplinary future research directions, as well as still-untapped resources for the design of innovative experiments and effective defensive mechanisms. This chapter is based on an article published in a refereed journal [59].

In Chapter 5 we answer RQIII by performing two simulated phishing campaigns against a European university and an international company operating in the consultancy sector. We derive different attack delivery techniques from the user notifications literature, and identify the relation between notification techniques and cognitive attacks (i.e., implementing persuasion principles) commonly employed in phishing. We ran our experiment targeting $n = 747$ employees across different roles within the respective organizations, and measure the relative efficacy of the adopted tailored phishing techniques. The study results show that overall employees are highly susceptible to tailored phishing attacks, with an attack success rate between 10% and 30% across user roles and organizations. The tailoring of the campaign to the victims (i.e., using target-related information) appears to be more important than the mere presence of a cognitive attack itself (i.e., the usage of persuasion techniques), whereas the adoption of notification methods can boost attack success (up to three times). The implications of this study shed light on the relations between well-known and novel phishing techniques across different organization types and employee categories whereby, for example, company employees and junior staff at the university were significantly more susceptible to authoritative persuasion methods than senior and support staff at the university. Finally, we report the velocity at which users fall for the attack across user category and organization type, which allows us to reason on potentially effective response practices. The study described in this chapter is published in the proceedings of a peer-reviewed conference [61].

In Chapter 6, we contribute to answer the second part of RQIII concerning additional mitigation strategies for phishing attacks that aim at stealing web credentials. We develop a novel approach to identify which website a phishing web page is imitating by means of both textual features extracted from the metadata of the page and visual features (regions) extracted from a screenshot of the page. The evaluation show that, compared to previous text-based classifiers, our method reduces the phishing misclassification rate by 67%, for an overall accuracy of 99.66% on our dataset. The tool, implemented as a browser extension, can support users in the detection of phishing websites, mitigating the threat of websites aiming to steal user credentials. This study enables a novel, integrated research line to investigate the complex interaction between users and semi-automated decision support tools. The studies described in this chapter are published in the proceedings of peer-reviewed conferences [58, 348].

In Chapter 7, we follow up to the second part of RQIII by observing how the untapped potential of crowd-sourced detection and reporting can assist with response to advanced phishing attacks. We do so by interviewing employees that reported the phishing emails of our tailored

campaign to the IT department of the mid-sized European university. The preliminary results uncover the respondents' inability to generalize the rationale for notifying a suspicious email. We provide pointers for future work to improve the phishing reporting process based on mental models of individuals that are arguably better predisposed to detect complex attacks. The study described in this chapter is published in the proceedings of a peer-reviewed conference [56].

In Chapter 8, we follow up on RQIV by investigating employees' reactions to a tailored phishing attack in small enterprise. We first run a field experiment to demonstrate how an attacker can leverage publicly available data in a tailored phishing attack. Subsequently, we interview nine employees to understand the cognitive processes underlying the detection and response to our campaign. Interestingly, we observe the engagement of a strong community reaction whereby employees take immediate action to protect each other from the attack. Our findings show that the identification of an inconsistent pattern played a central role in detecting our phishing campaign. This can potentially be attributed to the relatively small size of the company, where interpersonal familiarity is prevalent. The tailored nature of the attack prompted certain employees to swiftly inform the wider group, resulting in a defensive reaction surprisingly faster than what is typically expected at larger organizations. The study described in this chapter is published in the proceedings of a peer-reviewed conference [60].

In Chapter 9, we address the second part of RQIV by developing and testing a theoretical model that explains intention to report as a function of certain human factors, such as personality traits, attitudes towards the organization and co-workers, and beliefs. Our empirical evaluation shows that accounting for different types of human factors at both individual and organizational levels provides a more comprehensive understanding of their effects on individuals' positive cyber security behaviors and intention to report phishing emails. For example, the altruism trait has a significant positive impact on individual's intention to report phishing emails, whereas it has no impact on generic positive cyber-security behaviors. Surprisingly, the sportsmanship trait (those who tend to tolerate less-than-ideal situations, such as the nuance of reporting a phishing email) has a negative influence on reporting phishing. This carries a series of implications on both theoretical and practical levels and can help organizations to improve their overall security posture. The study reported in this chapter is published in the proceedings of peer-reviewed conference [222].

Chapter 10 revisits our research questions with the relative contributions, and provides brief discussion points for future research.

1.5. Publications

Our research led to the following publications in conference proceedings and journals:

1. **P. Burda, T. Chotza, L. Allodi, and N. Zannone**, *Testing the Effectiveness of Tailored Phishing Techniques in Industry and Academia: A Field Experiment*, In the 15th International Conference on Availability, Reliability and Security (ARES), ACM, 2020

2. **P. Burda, L. Allodi, and N. Zannone**, *Don't Forget the Human: a Crowdsourced Approach to Automate Response and Containment Against Spear Phishing Attacks*, In IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), IEEE, 2020

3. **P. Burda, L. Allodi, and N. Zannone**, *Dissecting Social Engineering Attacks Through the Lenses of Cognition*, In IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), IEEE, 2021

4. **B. van Dooremaal, P. Burda, L. Allodi, and N. Zannone**, *Combining Text and Visual Features to Improve the Identification of Cloned Webpages for Early Phishing Detection*, Tn the 16th International Conference on Availability, Reliability and Security (ARES), ACM, 2021

5. **P. Burda, L. Allodi, and N. Zannone**, *A Decision-Support Tool for Experimentation on Zero-Hour Phishing Detection*, In Foundations and Practice of Security (FPS), Springer LNCS, 2022

6. **P. Burda, A. Altawekji, L. Allodi, and N. Zannone**, *The Peculiar Case of Tailored Phishing against SMEs: Detection and Collective Defense Mechanisms at a Small IT Company*, In IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), IEEE, 2023

7. **I. A. Marin, P. Burda, N. Zannone, and L. Allodi**, *The Influence of Human Factors on the Intention to Report Phishing Emails*, In Conference on Human Factors in Computing Systems (CHI), ACM, 2023

8. **P. Burda, L. Allodi, and N. Zannone**, *Cognition in Social Engineering Empirical Research: a Systematic Literature Review*, In Transactions on Computer-Human Interaction (ToCHI), ACM, 2023

Other publications during the PhD that are not included in the dissertation:

1. **P. Burda, C. Boot, and L. Allodi**, *Characterizing the Redundancy of DarkWeb .onion Services*, In the 14th International Conference on Availability, Reliability and Security (ARES), ACM, 2019

2. **M. Campobasso, P. Burda, and L. Allodi**, *CARONTE: a Crawler for Adversarial Resources Over Non-Trusted, high-profile Environments*, In IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), IEEE, 2019

3. **L. Kersten, P. Burda, L. Allodi, and N. Zannone**, *Investigating the Effect of Phishing Believability on Phishing Reporting*, In IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), IEEE, 2022

2

Background

To better understand the following chapters, we introduce background notions on Social Engineering, including attack phases and tailoring, the cognitive processes relevant to SE, and various methods to counter phishing attacks.

2.1. Social Engineering and tailored phishing

The term 'Social Engineering' is used in information security to refer to a type of attack wherein an attacker manipulates individuals to compromise the confidentiality, integrity, and availability of data and processes by exploiting human vulnerabilities [368]. Figure 2.1 represents the main phases of an SE attack. In the initial phase, attackers define specific attack objectives (e.g., stealing credentials, obtaining sensitive information), identify potential target(s) of interest, and gather relevant information in the reconnaissance and intelligence phases [15]. The gathered intel can include contextual information on the targets and their environment to support attack orchestration and execution. During the orchestration phase, attack artefacts such as phishing emails and websites are crafted accounting for the available information. This includes forging decoy identities, constructing believable pretexts and *tailoring* the attack towards the target's environment, such as tuning the language to match the tone and syntax to which the target is accustomed to within that context. The adaptation of attack artefacts to their targets has recently become a prominent characteristic of SE attacks [15, 47], in stark contrast with 'classical' SE attacks that are untargeted in nature and employ simple techniques to persuade their victims [40, 108, 277].

The constructed artifacts are then delivered to the targets in the execution phase. Finally, the attacker waits that the targets execute the attack payload (e.g., they submit their credentials) for target exploitation. An attack can be cycled through multiple subsequent stages, during which the attacker can collect additional information about the victims and attack environment, and escalates from there (e.g., to move horizontally or vertically in an organization's structure from the current advantage point) until it reaches its final objective [15, 148].

Attackers can exploit a variety of communication channels to deliver their attacks, such as email (phishing), voice calls (vishing), SMSs (SMShing) or Social Networking Sites (SNS),

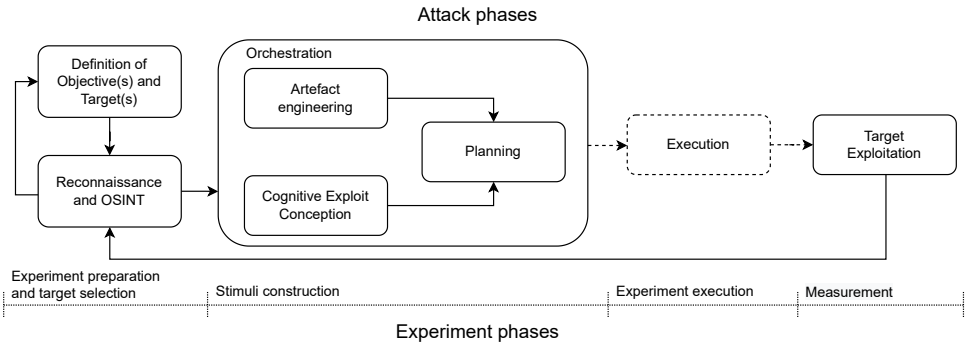


Figure 2.1: The schema of SE attack (top) and experiment (bottom) phases.

to lure targets and elicit information [339, 357]. Some techniques involve physical displacement, where the attackers physically perform parts of the orchestration and execution stages by infiltrating buildings or visiting locations of interest, to achieve their goals [55, 338, 354]. Examples are tailgating or dissemination of malicious QR codes, USB drives and decoy wireless access points. SE attacks also play a role in *Advanced Persistent Threats* (APTs), where the threat actors have access to nation-grade resources and carry out complex operations, such as (open source) intelligence and lateral movement, to engineer and deliver sophisticated artefacts, such as tailored phishing emails or USB drives with payloads triggering 0-day exploits [47, 198]. Experiments in SE often simulate the phases of real SE attacks (bottom in Figure 2.1).

Tailored phishing

SE attacks have historically been untargeted and typically aimed at harvesting credentials for payment or consumer services (e.g., phishing campaigns impersonating PayPal, eBay, etc.) by employing common persuasion techniques built on urgent and authoritative requests [277]. Recent reports highlight the emergence of highly tailored SE attacks, mostly spear and tailored phishing, in which information on the target and their context is adapted to the receiver, and sometimes diluted in multiple stages across different media [39, 40, 108, 352]. Spear phishing specifically targets small groups or single individuals and encompasses iterative information gathering and attack engineering, such as carefully personalized pretexts with compromised information or senders from compromised accounts [144, 220]. Tailored phishing differs from spear phishing, for example, in terms of lower attack sophistication but a higher potential for scalability. Similarly to generic phishing, tailored phishing is a single-stage attack and of a 'hit-or-miss' nature; however, it encompasses a more advanced reconnaissance and open source intelligence phase, similar to spear phishing, to gather additional information about the victims and their organization (e.g., name, organization's domain name, communication processes and practices at the organization). An example of tailored phishing can be drawn from a phishing exercise at a US military academy where the authors implemented a widely known communication practice at the academy in the email salutation [113]. Listing 2.1 reports the phishing email sent to the 512 participants.

Listing 2.1: Tailored phishing attack from [113].

```
From: sr1770@usma.edu [mailto:sr1770@usma.edu]
Sent: Tuesday, June 22, 2004 4:57 PM
To: cadet@usma.edu
Subject: Grade Report Problem

There was a problem with your last grade report.

You need to: Select this link Grade Report and follow the
instructions to make sure that your information is correct;
and report any problems to me.

Robert Melville
COL, USCC
sr1770@usma.edu
Washington Hall, 7th Floor, Room 7206
```

The campaign deployment was timed a few weeks before end of semester (end of June), when emails regarding final exams can be a relevant topic for the targeted cadets. With just two easily obtainable elements (the email sign-off, 'COL', and timing) the authors of the experiment were able to register clicks from 80% of the 512 participants. To illustrate the possible 'tailoring degrees' of the phishing scenario in [113], Table 2.1 presents potential variations in targeting and sophistication of email features commonly forged in phishing attacks[1]. Each feature of Table 2.1 can be crafted to appear more relevant and realistic to the target (from generic to spear-phishing) by gathering additional target-related information (e.g., reconnaissance, OSINT) and by engineering more advanced attack techniques and procedures (e.g., spoofing the sender address[2], detection evasion techniques). What distinguishes tailored and a spear-phishing in the example of Table 2.1, is the size of the potential target base: from any user at the academy (generic), the wide group of cadets (tailored), and to a specific cohort of cadets (spear). A second distinction is the quality and quantity of information needed to make the attack relevant to the targeted group: the tailored and spear variants both attempt to match the target context by means of the sender, timing, pretext and signature; the spear variant, however, employs qualitatively and quantitatively different bits of information, such as precise sender name, address, course name and signature, and more advanced attack techniques, i.e., spoofing. Importantly, such considerations can be extended to the following attack stages (i.e., the payload landing page) and to other attack vectors (e.g., text messages). The expected impact on attack success of the three variants likely differs as well.

[1] Other email features might be relevant as well, such as the email's look and feel (e.g., graphical and layout features). As the attack in [113] does not rely on look and feel features, we omit this feature in our example in Table 2.1.

[2] At the time of the example in [113], address spoofing was more relevant and security awareness was lower than today. This likely influenced the high attack success rate. Currently, spoofing techniques are less effective.

Table 2.1: Examples of possible 'targetization degrees' of the phishing email in [113].

Feature	Generic	Tailored	Spear	Description
Sender	danygedf@yahoo.com	sr1770@usna.edu	real.surname.col@usma.edu	The address is increasingly more tailored to the target and implementation sophisticated: more realistic domain (typo-squat and spoofing) and username (name/surname). The tailored variant is potentially relevant to any student at the academy, whereas the spear variant might be mostly relevant for students familiar with the sender.
Timing	Saturday, December 25, 2004 3:00 AM	Tuesday, June 22, 2004 4:57 PM	Tuesday, June 22, 2004 4:57 PM	Whereas timing in the generic variant can be unrelated to the actual pretext/context, in the targeted variants the timing is critical.
Subject	Action Required	Grade Report Problem	[Kinesiology] Grade Report Problem	The subject may attempt to convey more relevant and/or reasonable preview of the contents, as a function of attacker assumptions/knowledge. The tailored variant is applicable to any student, while the spear one is relevant for kinesiology attendants (on top of familiarity with the sender).
Body/ pretext	Dear user, the web-mail information system has been updated. Please follow this *link* to make sure your account is up to date. Your record is set to expire in 3 days. Update NOW to avoid the deletion of the account.	There was a problem with your last grade report. You need to: Select this link Grade Report and follow the instructions to make sure that your information is correct; and report any problems to me.	There was a problem with your last grade report in Kinesiology course (N. D114). You need to: Confirm your Grade Report on D114 report page and follow the instructions to make sure that your information is correct; and report any problems to me.	The body contains the 'core' of the attack: the pretext and the attack payload (URL). It may include information relevant to any student at the academy (generic), to students that attended an exam (tailored) or to the group that attended the kinesiology exam (spear). Additional elements that might increase the tailoring and relevance are the terms 'COL' (both tailored and spear) and 'D114 report page' (spear only).
Sign-off	IT desk & services USCC	Robert Melville COL, USCC sr1770@usma.edu Washington Hall, 7th Floor, Room 7206	Actual Real Name COL, USCC Washington Hall, 2nd Floor, Real Room 2206	The signature of the spear-phishing variant may include truthful information, as opposed to just realistic or loosely targeted in the tailored and generic variants.

More recent instances of tailored phishing in the wild concern organizations' credential harvesting campaigns where attackers integrate specific information about targets in meaningful ways, including: details within the subject of the email relevant for some segment of business operations; 'login' pages where the user's email or name is already filled in; and custom building of 'login' templates to match the look and feel of the corporate email systems used by the specific companies they are targeting [326][3]. The information used in tailored attacks can be gathered and/or implemented at a larger scale (than in spear phishing) with automated tools (e.g., [110]) and used to craft believable phishing artifacts, providing a higher level of sophistication than generic attacks, with relatively low effort (e.g., [281]). Therefore, tailored attacks may yield a similar impact to spear-like attacks but at a larger scale thanks to a lower overall attack effort or cost.

Regardless of the level of tailoring and sophistication, SE attacks tend to exploit 'vulnerabilities' inborn in human cognition, e.g. faulty beliefs and cognitive patterns [62, 114, 174, 303, 332]. Attackers engineer *cognitive attacks* by constructing artefacts able to exploit target's processing weaknesses with the aim of convincing their target to comply with their request (opening an attachment, a URL, or input confidential information to an attacker-controlled system). Therefore, the investigation of cognitive effects (such as the effects of persuasion techniques on the outcome of a phishing attack [379]) and the involved processes (such as priming subjects before deploying an attack [174]), must be considered to understand the underpinning mechanisms that bring to victimization. In the next section, we introduce the preliminary concepts on human cognition that are relevant for the following chapters.

2.2. Cognition in Social Engineering

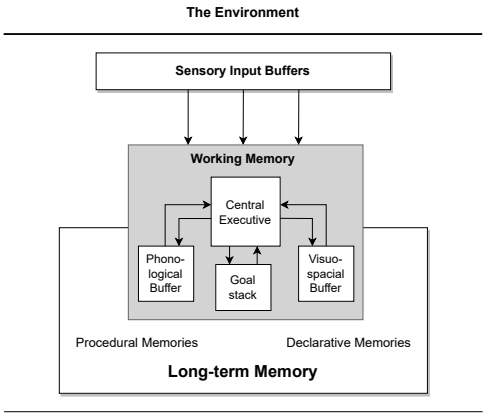


Figure 2.2: The human information processing system according to the Working Memory model (arrows mean flow of information and control) [146].

The term "cognition" refers to the abstract information processing that is implemented by neurons in the brain [280, p. 112]. Cognitive sciences have identified a general set of compo-

[3]Further examples of tailored phishing are the topic of Chapters 5 and 8.

nents that constitute the architecture of human cognitive processes, particularly in the fields of psychology, linguistics, and neuroscience [27, 146]. This results in a variety of theories and models accounting for mental capabilities of perception, memory, attention and reasoning among others. For example, the Working Memory model (shown in Figure 2.2) provides the currently dominant view of how conscious thinking occurs in the mind [36]. In this theory, a set of limited capacity systems store and manipulate information incoming from sensory systems and (long-term) memory. An attentional control system feeds, translates and retrieves temporarily stored (visual and phonological) computations in a short-term memory by prioritizing certain information over short periods of time [34]. The results of these manipulations can interact with other conscious and unconscious brain systems, such as accessing the long-term memory or functions responsible for movement.

Overall, perspectives in cognitive sciences come in various forms, ranging from detailed mathematical or computational models of specific phenomena, to broad theoretical frameworks that try to explain a wide range of phenomena [36]. In this thesis, we are interested in the latter – high-level – representation of cognitive processes that can be affected during an SE attack, as reported by various studies in this domain [62, 114, 174, 303, 332]. However, attempts to explore and characterize SE attacks through the lenses of cognition are so far unstructured and limited in number. Research in SE often focuses on single domains, such as technological, human-related or design-related, with few attempts that bridge various domains. Following, we provide an overview of related work on human cognition in SE.

Human cognition in SE

SE research focused on human cognition is relatively sparse and heterogeneous. For example, Cranor [84] proposes a framework for reasoning about the human in the loop to analyze the root cause of security failures attributed to human error. Their work is based on the Communication-Human Information Processing (C-HIP) model from warning science literature [83]. This framework illustrates the processing of information by a human receiver whose behavior is dependent on a set of processing steps, personal characteristics and environmental disturbances. However, the scope of this work is to facilitate the design and analysis of secure systems that rely on humans, such as supporting users with an anti-phishing tool, without contextualizing SE attacks therein. The survey by Pfleeger and Caputo [276] reviews behavioral science findings relevant to cyber-security, which partially cover features of cognitive processes, such as information elaboration and behavior; yet, no SE attacks are characterized and related to the resulting insights. Montanez et al. [238] map a selection of SE attacks features into a basic and selective framework of human cognition functions, including perception, memory, decision-making and behavior among others. The authors advocate to treat SE attacks as psychological attacks by extending the standard framework of human cognition to accommodate SE attacks, which is closely related to the goals of this thesis. They discuss the role of short and long-term cognitive factors (e.g., workload, experience, etc.), memory and attacker effort, and their effects on persuasion of the targets; however, no details on how to instrument their proposal to characterize SE attacks are provided. Steinmetz et al. [321] interview social engineers (attackers) on the process and attributes of SE attacks, and reveal that SE deceptions are intractably intertwined in situational, cultural,

and structural circumstances. This work overlaps with this thesis in the intention to understand the fundamentals of SE attacks' success, but only from an inherently social psychology perspective with no specific attack characterization.

Characterization of SE attacks

Research focused on the attack side of our problem is more available and, generally, comes together with a characterization of defenses against SE. For instance, Heartfield and Loukas [147] propose a taxonomy of semantic SE attacks along with their characteristics and a review defense techniques, similarly to the work of Salahdine and Kaabouch [303] and Purkait [286]. Darwish et al. [85] investigate the relationship between victims' characteristics, such as demographics and personality traits, and phishing attacks, along with some detection techniques. Tetri and Vourinen [332] introduce a conceptual framework for SE that touches attack characteristics, a few target and setting-related factors, and the execution SE attacks. Sommestand and Karlzen [320] review phishing field experiments by looking at experimental variables, results (susceptibility rates) and experiment design features, such as hypotheses, control variables, etc. Most of these works on SE attacks (and defenses) do not treat cognitive-related aspects [85, 147, 286, 303]. The remaining works [320, 332] relate attack characteristics and certain cognitive processes, although only implicitly and partially.

For the purpose of this thesis, we adopt the perspective of the cognitive framework for SE proposed in Chapter 3. This framework is grounded on existing, well-established theories of cognition mapped to the SE domain, and is, therefore, suitable to structure and analyze SE attacks from a cognitive perspective. The development of the framework, its usage and relation with the broader literature on human cognition are the topic of Chapter 3.

2.3. Measures to counter tailored phishing

The most investigated SE attack in research is undoubtedly phishing, with a particular focus on the design of technical countermeasures based on block-listing and machine learning-based detection [11, 320]. Research on tailored and spear-like phishing – that is, targeted and sophisticated variants of phishing – primarily investigates attack and training effectiveness. Despite these efforts, no definitive solution to phishing, more so to spear-like phishing, has been found yet. Indeed, the reported mean susceptibility rate to phishing attacks across various experiments and measurements is 21% [320], while spear-like phishing exhibits more impressive numbers: for example, Ferguson and Bargh report that 80% cadets in a military academy were successfully phished in a training exercise (see Section 2.1) [113]; Kumaraguru et al. successfully phished around 50% of the subjects in their experiment [192]. In their *context aware* phishing campaign against Indiana University students, Jagatic et al. obtained a success rate of 70% [161]; Caputo et al. report a spear-phishing susceptibility rate of 60% in their first trial [66]; Burns et al. obtain a click rate of 70% [62].

The high susceptibility rates achieved by sophisticated variants of phishing indicate that current countermeasures might not be well-suited against this type of attacks [15]. Next, we discuss the various methods to counter phishing attacks and their effectiveness against spear-like phishing, starting from solutions employed to reduce phishing susceptibility of potential

targets to response strategies such as infrastructure take-downs.

Prevention

The ideal remedy to spear-like phishing and, in general, to phishing is to make potential targets immune to the attack altogether. Preventive measures typically encompass training users to recognize specific attack features and rising their awareness of the threat. Prior studies show that training has a significant effect in reducing generic phishing susceptibility, albeit leaving margins of untreatable portions of subjects around 10-15%, even with repeated training [194, 369]. The same effects, however, might not be achieved against spear-phishing. Kumaraguru et al. performed a controlled field experiment to test the effectiveness of training tailored to spear-phishing, showing no significant differences between generic and spear-phishing training effects [194]. Caputo et al. report no effects of training (and awareness) at all when conducting a spear-phishing attack in their experiment [66]. Burns et al. report a marginally significant effect of training tailored to the detection of spear-phishing attacks, reducing phishing susceptibility rate from 70% to 54% after five weeks of training [62]. Other works show that training effectiveness decreases over time [31, 52]. Even if some reduction can be achieved, the underlying problem of the training and awareness campaign lies in the fact that spear-like attacks can take very different forms, making the attack difficult to be recognized by users, and requires much less victims than generic phishing to achieve the desired objectives [66]. For example, the 'ideal' spear variant of Table 2.1 would be virtually indistinguishable from a legitimate message and somewhat difficult to be trained for. As some users will still remain vulnerable, training and, in general, preventive measures alone may not minimize the attack surface enough to neutralize or effectively contain sophisticated attacks.

Detection

The most popular approach to the detection of phishing attacks is *artifact filtering* in an anti-spam fashion, including emails, URLs and attachments [77, 101, 149, 180, 327, 389]. These countermeasures have been implemented using numerous methods, such as data mining, machine learning, heuristics and allow/block-listing [116]. Solutions based on machine learning techniques might be affected by a large number of false positives and require continuous retraining [77]; furthermore, they also are not generalizable across domains [347]. A few studies show how these solutions can be bypassed, for instance by legitimizing the sender (via multiple iterations) to appear less "anomalous" to an anomaly detection system [77] or by taking over a legitimate account, e.g. one of the target's secondary accounts or one of their associates [148]. Another body of research focuses on the examination of phishing sites and server characteristics, and relies on block-listing. Some of these works leverage crowd-sourcing [117, 241] and reputation systems [212] to improve accuracy and speed. While these solutions have proven to be suitable against general phishing and known threats, they face significant limitations against tailored and spear-phishing, as block-lists do not generalize well to unknown attack instances (new URLs, clean IPs, low email volume, etc.) [152]. The fundamental drawback of automatic detection techniques against spear-like attacks is the unforeseeable nature of attack characteristics and artifacts, such as pretexts and links [15, 47]. Such artifacts are meticulously crafted to fit targets' context (demograph-

ics, work and previous social interactions) [47], and to fly under-the-radar by employing legitimate-but-compromised or vanilla websites and by targeting a small numbers of recipients, such as the tailored and spear variants of Table 2.1 [77].

Among automatic detection methods, some approaches are able to detect unknown attacks by relying on the visual similarity of certain attack features. This is the case of phishing websites that are often the payload in a phishing email. These techniques use features such as the logo, the screenshot of the webpage or other features to compare two websites and determine which website a phishing web page is imitating [6, 72, 210, 213]. However, their ability to detect phishing attacks depends on their ability to find the impersonated legitimate website [348]. This important limitation is evident in the state-of-the-art [210, 213] where it is often addressed by narrowing the scope to a predetermined target list of sites or brands that covers specific classes of phishing attacks, potentially, leaving out less popular brands or organizations.

As a last line of defense, detection can also be accomplished by security analysts, who often are superior to automatic tools. However, this solution requires that artifacts of interest are first reported to security analysts by the targets themselves. Thus, these countermeasures leave advanced attacks to remain undetected or to be detected too late, when the attack may have already propagated.

Response

Response strategies are typically employed by an organization's IT department and security operation centers (SOC) to mitigate the damage of an attack when it occurs. Response teams are aided by both detection techniques outlined above and notifications from users that detect and report the attacks [188]. A response can be initiated immediately after detection (e.g., employees notifications) or later after the attack effects have manifested (e.g., a data leak was identified) [350]. In the former case, incident reporting can alert interested parties (e.g., subsidiaries or clients) of the incoming threat. The remediation procedure can combine attack interception by blocking traffic or investigating rogue domains, although attacks are typically characterized by a short duration, thus hindering such attempts [47]. In particular, these containment procedures are sensitive to time delays [152, 352], especially in case of unknown attacks [188]. Prior studies have shown that the response time often does not match the velocity of spear-phishing attacks where the expected compromise and exfiltration timelines are in the order of minutes and, at most, hours, while discovery and containment are in the order of hours or days [350, 352]. For instance, previous work reports a 50% success rate after 6 hours from the launch of the attack [161], whereas the same rate was achieved within only 2 hours in [192]. Recent measurements show that the typical 'mass' phishing campaign lasts on average 21 hours, with half of the victims after 8 hours and detection only at least in 9 hours on average [255]. Overall, there are very few studies that investigated the velocity of attack propagation in the context of phishing, with only one such study demonstrating the viability of a timely response to spear-like attacks in organizational setting [196].

I

Characterizing the human attack surface

3

Breaking down SE complexity: dissecting SE attacks through the lenses of cognition

Social engineering (SE) attacks leverage the fundamental weaknesses ingrained in human thinking processes, enabling malicious actors to manipulate system users to perform actions against their own self-interest. Research in SE explores the complex relationship between the technical aspects of an attack and the cognitive factors that define its human element. However, the multidisciplinary nature of the SE domain presents challenges in identifying gaps and open research questions, and interpreting experimental outcomes. As a result, it becomes difficult to reconcile different findings and reach a comprehensive understanding of the 'SE attack surface'. For this reason, in this chapter we present, showcase, and analyze a framework to dissect SE attacks (see RQI). The framework is meant as an instrument to structure and analyze SE attacks of varying sophistication, isolating specific features and their effects at the cognitive level, and providing a common structure for comparisons across different attacks. We showcase the framework against attacks which main characteristic is their sophistication and increased tailoring towards the targets. The analyzed attacks are two simulated phishing attacks from the literature and two real, highly-targeted SE attacks reported in the wild. We carry out the analysis by isolating and relating effects and techniques adopted by the attackers to the target's cognitive process. The proposed framework, therefore, has the potential to enable a new way to analyze SE attacks from the literature, which will be essential, in Chapter 4, to identify open gaps in empirical SE research.

This chapter is originally published as P. Burda, L. Allodi, and N. Zannone, "Dissecting Social Engineering Attacks Through the Lenses of Cognition", In *IEEE European Symposium on Security and Privacy Workshops* (EuroS&PW), IEEE, 2021, pp. 149–160

3.1. Introduction

In Chapter 1, we have seen that SE attacks are cognitive attacks aiming at deceiving individuals by exploiting 'vulnerabilities' inborn in human cognition, with the goal of gaining access to confidential information and/or deliver malware on the target's system [233]. The last years have witnessed an increased sophistication of human-based exploitation techniques, evolving from less sophisticated *'your email account is full, click here to reset your password'* type of attacks to tailored and well targeted attacks exploiting target information [47]. Research in this area has, therefore, taken a multidisciplinary approach to grasp the nuances of the interactions between the technical aspects of an attack, and the cognitive aspects characterizing its human element. However, empirical research results are often contrasting and hard to contextualize, making it hard to derive effective defensive measures for real world applications. The question of how to develop SE research to support a coherent interpretation of cognitive effects, and test related countermeasures, remains an open problem.

In Section 1.2, we argued that the primary reason for this is the lack of a structured, shared understanding of the dimensionality of the SE problem within the research community. For example, studies often focus on single domains (e.g., technological, human-related or design), but experimental designs capable of isolating interaction effects across domains are hard to devise without a clear framework of the underlying cognitive processes. Similarly, most empirical research results are limited to 'untargeted' attack scenarios, whereas little understanding remains of the nuances of *targeted* attacks and exploitation of target-relevant information (e.g., memories). On the other hand, targeted attacks are becoming increasingly relevant to the overall threat landscape [15, 47], stressing the importance of filling the gap between SE research results and real-world situations.

In this chapter, we underline that understanding the cognitive processes involved in a SE attack is fundamental to (a) advance the field of empirical and theoretical research in SE by identifying gaps and effect interactions; (b) provide a framework to evaluate and contextualize research results; (c) characterize the SE attack surface to, for example, be able to measure threat levels, or devise research toward effective policies to thwart targeted SE attacks. To this end, we develop and showcase a cognitive framework for characterizing SE attacks based on theories and models of human cognition drawn from the field of cognitive sciences. The framework can support the design of experiments in the SE domain (e.g., by providing a structure to isolate cognitive effects), as well as being employed to characterize and study existing, sophisticated attacks in the wild, thus, helping uncovering novel attack techniques whose effects may be tested in experimental settings. To showcase the framework's application, we analyze two academic experiments simulating phishing attacks [61, 256] and two real, highly tailored SE attack cases [15, 47] to illustrate how the framework can be used to identify gaps and ways forward.

The chapter is structured as follows: Section 3.2 describes the framework derivation from extant theories of cognition and the framewrok itself. In Section 3.3, we illustrate the framework usage and applicability to four SE scenarios (two simulated attack experiments and two real SE attack cases). We discuss implications for research and practice of the proposed framework in Section 3.4. Finally, Section 3.5, provides a detailed summary and comparison with other similar works.

3.2. Cognitive framework of SE attacks

3.2.1. Framework derivation

Cognitive sciences identify a general set of components that constitute the architecture of the cognitive processes of the human mind whose body of evidence stems from the fields of psychology, linguistics and neuroscience [27, 146].

To derive the building blocks of our framework, we investigated popular theories of cognition in the cognitive sciences [27, 146], and mapped those to the SE domain. Table 3.1 presents an overview of the extant theories and models of cognition summarizing the common perspectives of cognitive processes. We examine the cognitive processes corresponding to human information processing that can be affected by an attacker during an SE attack, as reported by various studies in this domain [62, 114, 174, 303, 332]. The identified 'building blocks' of the human cognitive processes that are relevant to SE attacks are reported in Table 3.2, pictured in Figure 3.1, and presented in detail in the next section.

3.2.2. Framework building blocks

Stimulus: The stimulus is any input (e.g., an event, a sound, a message) that triggers a cognitive process. In the SE context, the stimulus represents the means by which the attack is delivered to its (human) target. A stimulus is characterized by attributes describing its content and form. Examples of attributes can be presence/absence of a spoofed address in an email [221], style of writing [215], or the presence of text aimed to evoke past memories of the target [380]. These attributes contribute in determining which components of the framework may be "activated" as the information is processed.

Parameters: Parameters are used to capture contextual information during the cognitive process. Context is assumed to influence many aspects of the production and understanding of text and speech, and is defined as the set of subjective constructions or "definitions" of the *relevant dimensions* (i.e., parameters) of social or communicative situations [89]. We distinguish between *attack parameters* and *target parameters*. Attack parameters represent the assumptions that the attacker makes on the targets and their context. Target parameters characterize the properties of the target and the context in which the target is when the external stimulus arrives. Thus, target parameters mediate the processing pipeline from stimulus to behavior and define the overall context in which the cognitive processes takes place. As shown later, this distinction allows us to reason on the level of targetization of an attack and its effectiveness as the success of the attack is strongly related to the alignment of attack parameters with target parameters [125, 132].

Perception: Perception decodes the sensory information from an incoming stimulus. Perception is a complex process spanning from audio-visual interpretation to features integration and pattern recognition. In particular, perception functions as a signal receiver that translates the stimulus into a '*percept*' and automatically loads, from Long-Term Memory (`LTM`) further associations, experiences and judgments related to the stimulus and its attributes based on the contextual parameters of the subject [27, 146]. Perception fetches this information and makes them readily available for the upcoming computation, similarly to what

Table 3.1: Theories and models extracted from the cognitive science literature.

Theory/Model	References	Key aspects
Working Memory model (WMM)	Baddeley [36], Hastie and Dawes [146], Anderson (pp. 129-131) [27]	This model is a multi-module model where input and output modules encode information from sensory systems. The main module is the `Working Memory` (`WE`) where manipulation of information from *perception* modules occur. A major system is the `Long-Term Memory` (`LTM`) that contains all sorts of information including procedures for thinking and deciding. The controller of these modules is the `Central Executive` (`CE`) which functions as the *attention* selector and controller of explicit and implicit cognition.
Dual-process theories (DpT)	Evans [111] Sanfey et al. [305] De Neys and Glumicic [86] Oppenheimer [259], Hastie and Dawes (pp. 21-27) [146]	Dual-processing models theorize two different modes for cognitive processing, one highly cognitive-intensive, and the other engaging only low-cognition. A common conceptualization of these modes are "System-1 and System-2", where two systems compete for over response: one is unconscious, rapid, automatic, and high capacity; the other is conscious, slow and deliberative. Regardless of the conceptualization, these theories suggest a mixed use of fast and approximate computations based on *heuristics* where only the final product (e.g., behavior) is posted in consciousness. The engagement of higher or lower cognition is mediated by exogenous and endogenous contextual variables, for example, the environment, fluency, etc.
Global Workspace theory (GWT)	Baars [32] Dehaene and Naccache [87] Baars and Franklin [33]	This theory focuses on conscious processing whose coordination and control depend on the `CE` and its access to a global workspace where other processes are running automatically and unconsciously. The cognitive cycle starts with a competition for consciousness of signals from *perception* modules and `LTM`, which then can proceed to the `WM` (via *attention*) if goal-relevant or picked up by other modules of the workspace. *Attention* modulates the access to consciousness. `WM` and `LTM` function as a substrate for the cognitive cycle.
Expected Utility theory (EUT)	Hastie and Dawes (pp. 551-552) [146]	This theory describes a rational decision making method under uncertainty where individuals seek the highest combination of subjective value (utility) and the highest (expected) probability of events. This decision-making technique can be enabled only by perfectly rational agents.
Prospect theory (PT)	Hastie and Dawes (pp. 655-658) [146]	This theory describes a decision making method under uncertainty where individuals seek the highest utility and the highest prospect (the potential to happen in a desired way) of events. The difference with the expected utility theory lies in the asymmetry of weighting the probability of events for which, for example, losses can have bigger weights than equal gains. This theory corroborates the dual-processing nature in dual-process theories.
Load theory of attention (LToA)	Lavie [201]	*Perception* processes all stimuli in an automatic mandatory fashion until capacity permits. In case of high perceptual load it is less easy to get distracted by goal-irrelevant stimuli. Whereas in high cognitive load it is easier to get distracted by goal-irrelevant stimuli. Thus, an effortful goal-relevant *attention* is necessary in directing executive functions (`CE`) and distractors can more easily disrupt goal-relevant processing.
Feature integration theory (FiT)	Anderson (pp. 62-65) [27]	People typically focus their visual attention on a stimulus before they can synthesize its features into a pattern. This happens in the *perception* step where early perceptual processing occurs and patterns are recognized. It follows that anomalies are easier to spot when their features do not mix well in a perceived pattern. Thus, selective *attention* is needed to perform an array search between similar features which is a more difficult task.

Table 3.2: Overview of the building blocks for cognition and social engineering.

Component	Theory	Description	Relevance for SE
Stimuli	All	Any event or object that stimulates the senses.	The stimulus represents the means by which the attack is delivered to the target, e.g., an email or a voice call. Its attributes can be presence/absence of a spoofed address in an email [221], style of writing [215], or the presence of text aimed to evoke past memories of the target [380].
Perception	GWT LToA FiT	A signal receiver that translates the stimulus into percepts. It is mediated by other cognitive processes, like `LTM` associations, that can be concepts, procedures and categorizations, e.g., facial features.	Before the (attack) stimuli arrive, the target may receive 'priming' stimuli that do not necessarily result in behaviour (hence are not represented explicitly in Figure 3.1) but may have strong effects on the subject's subsequent decisions [75, 174]
Attention	WMM DpT GWT LToA FiT	A set of systems that modulate the access to consciousness. It has a limited capacity whose allocation can be *exogenous* (controlled by the stimulus) or *endogenous* (goal oriented by the `Central Executive`).	SE attacks exploit the lower amount of attention payed to stimuli that may be of less relevance to a subject in a given moment but still calling for action, like urgent or authoritative requests. Such is the case with exogenous attention (generated externally by stimuli properties), which has been demonstrated to lead to higher deception rates [242].
Elaboration	GWT DpT PT EUT	A block responsible for reasoning, like making a decision. It evaluates the available information from the loaded percepts and memory. It allocates cognitive resources, e.g. `WM` or `Attention`, based on currents needs.	Its operation is influenced by many modulating factors, as personality traits or past experiences. Two known factors that significantly influence processing and, consequently, behavior in the context of SE are heuristics and anomalies [114, 145, 361].
Anomaly	DpT	A condition when *Elaboration* cannot deal with the computation due to, e.g., wrong or lack of contextual cues, and engages in effortful processing, like consciously directing attention and making use of `WM`.	This mechanism is employed in anti-phishing training to allow for anomalies to be triggered, e.g., a mismatch between URL and the expected domain name, where relevant (or "mediating") knowledge is instilled (e.g. what is phishing, what the URL means, etc) and applied in practice (e.g., embedded phishing exercise) [193].
Heuristic	DpT	A condition in which *Elaboration* block has found a satisficing rule and engages in low effort processing by relying on heuristics to evaluate information and make inferences.	The effects of heuristics are commonly exploited in all sorts of SE attacks, like phishing [375] or social networks [359], and are thought to significantly affect the success of attacks [61, 256, 379].
Behavior	All	The output of the process. It is the response of the whole system to the stimuli, like complying or not complying with the request in the stimulus.	Depending on their objectives, SE attacks and simulations can elicit different types of behavior which lead to different consequences. For example, the success of an attack can be measured just by clicks on links in an email or by submissions of credentials on bogus websites.
Parameters	—	Properties characterizing the context in which the cognitive process occurs.	The distinction between attack and target parameters allows us to reason on the level of targetization of an attack and its effectiveness as the success of the attack is strongly related to the alignment of attack parameters with target parameters [125, 132].

Substrate		Description
`Long-Term Memory`	WMM GWT FiT	A memory system where knowledge is held indefinitely. The two main types of memories are stored therein: explicit recollections of factual information and implicit procedural memories.
`Working Memory`	WMM GWT LToA	A limited capacity system allowing the temporary storage (Short-Term Memory) and manipulation of information necessary for complex tasks as comprehension, learning and reasoning.
`Central Executive`	WMM GWT	An attentional control system that voluntarily manipulates the `WM` functions.

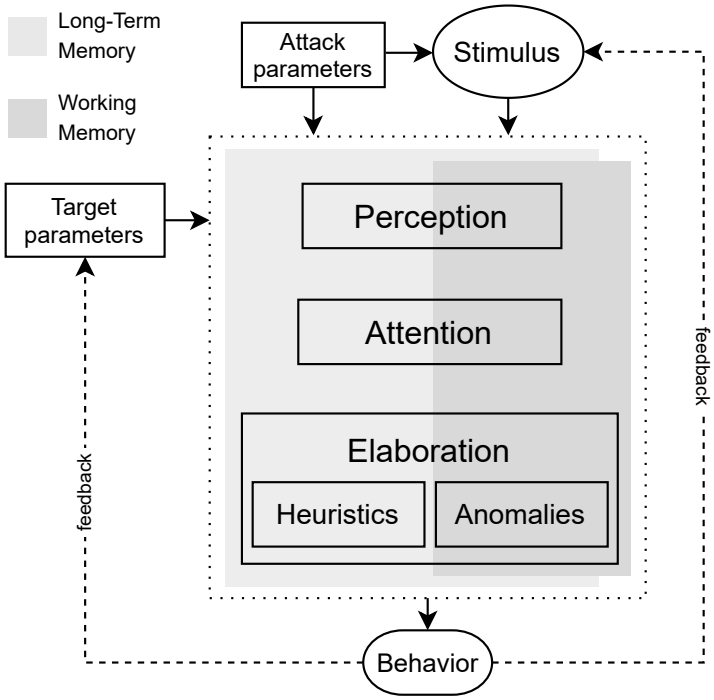


Figure 3.1: Generic framework of cognition for SE attacks

caching does in computing. In the SE context, perception is relevant with respect to SE attacks relying on priming [174]: before the (attack) stimuli arrive, the target may receive 'priming' stimuli that do not necessarily result in behaviour (hence are not represented explicitly in Figure 3.1) but may have strong effects on the subject's subsequent decisions [75]. Related concepts and information are stored in LTM, that could then be recalled in form of precepts when a new stimulus arrives, potentially conditioning the targets' decisions as detailed above.

Attention: Attention readies the central nervous system to process and respond to stimuli [27, 33, 201]. Attentional systems select information to process at serial bottlenecks and progressively filter irrelevant signals and information [27] (pp. 53-54). Within the contemporary theories of attention, two general models exist on different cognition levels [201] [27] (p. 54-74): '*peripheral attention*' relates to the sensory experience related to visual and auditory signals, whereas '*central attention*' relates to the semantics of the stimulus at a higher level of abstraction. Since we are concerned with higher level processing, i.e. when stimuli have already been pre-processed, we here consider '*central attention*', whose purpose is to select 'lines of thought' and to focus on a task while allowing for interruption by secondary tasks [27] (pp. 69-72). Central attention influences which and how stimuli are processed depending on the current set of goals in a given moment [217]. SE attacks exploit the lower amount of attention payed to stimuli that may be of less relevance to a subject in a given moment but still calling for action, like urgent or authoritative requests. Such is the case with exogenous attention (generated externally by the physical properties of stimuli), which has been demonstrated to lead to higher deception rates in the study of Morgan et al. [242] where they set participants' (central) attention to be endogenous (engaged explicitly by internal goals) or exogenous to react to malicious pop-ups.

Elaboration: Elaboration is responsible for processing the information incoming from the other blocks and information stored in memory. The processing involves various conscious and unconscious mental operations performed by a multitude of interconnected and distributed sub-modules [27, 87, 146]. As we are concerned with the cognitive features that can affect cognitive functions with respect to SE attacks, functional and neurobiological definitions of such sub-modules, or the mapping of psychological modules with specific neural circuits, is not of relevance here. Within the scope of this work, the elaboration block is therefore treated as a black box whose operation is influenced by two modulating factors that are known to influence processing and, consequently, behavior in the context of SE: heuristics and anomalies [114, 145, 361].

Heuristics are fast and implicit (that is, not available to introspection) psychological rules that aid judgment and decision making in the elaboration phase [146]. Their use is akin to 'speculative execution' in computing where heuristic processing employs a number of cognitive shortcuts that lead to appropriate behaviour under most circumstances. Heuristics can emerge from the need of having adequate but fast decisions (e.g., triggering innate behaviour under life-threatening situations), or to lower the cognitive burden associated with repetitive, pattern-specific decisions (e.g., breaking under a red light while driving, or to perform repetitive tasks) [175, 176, 340]. Cognitive biases such as those described by Cialdini, and often employed in SE research, can also be described as heuristics [73, 114]. Heuris-

tics are stored in Long-Term Memory and are mostly automatic and unconscious in nature [146]. The effects of heuristics are commonly exploited in all sorts of SE attacks, such as phishing [375] or social networks [359], and are thought to significantly affect the success of attacks [61, 256, 379].

The second influencing factor are *Anomalies*, anomalous conditions that take place when elaboration is unable to handle information that does not fit an automated processing pattern [175, 176] (e.g., a mismatch between URL and the expected domain name). Therefore, the CE (see Table 3.2) has to allocate cognitive resources to accomplish the current task, such as consciously directing attention to the processing of the anomaly, effectively creating a new goal for the elaboration. This requires employing the WM to handle the current task and reason on a judgment or decision by means of a wider set cognitive capabilities [87], for example making connections between experiences and knowledge stored in LTM with the current case [36]. This mechanism is employed, for example, in anti-phishing training to allow for anomalies to be triggered, where relevant (or "mediating") knowledge is instilled (e.g., what is phishing, what the URL means, etc.) and applied in practice (e.g., embedded phishing exercise) [193]. The availability of relevant knowledge (e.g., expertise) [364], the lack of cognitive resources (e.g., workload, stress) [242] or habits (e.g., context, personality) [361] are all exemplar factors that can condition the triggering of anomalies.

Behavior: Behavior is the output of the cognitive process (e.g., the decision to click a link). The last behavior can serve as a new stimulus and initiate a new cognitive cycle.

Substrate: The substrate represents the computational architecture on top of which the building blocks run [146]. The main components are described at the bottom of Table 3.2. The cognitive framework operates on a substrate made from the Long-Term Memory (LTM) and Working Memory (shaded in Fig. 3.1). These two components comprise multiple processing and memory sub-modules and constitute a 'workbench' for mental processes [33, 146]. The Central Executive (CE, not present in the figure for readability) is responsible for the coordination of mental processes, control of selective endogenous attention and inhibition of automatic responses [33].

3.3. Cognitive analysis of SE attacks

To illustrate the framework and its applicability to a range of SE scenarios, we apply it to model two SE attacks simulated in academic experiments [61, 256][1] and two real SE attack cases from the literature [15, 47] of varying 'sophistication'. This illustration showcases how both *real* and *synthetic* SE attacks can be interpreted and broken down using the proposed cognitive framework, similarly to the what is done in [84]. The symbols used through the description of SE attacks are shown in Table 3.3.

3.3.1. Breakdown of two SE academic experiments

[1]The example in [61] refers to the work carried out in Chapter 5.

Table 3.3: Notation

Symbol	Description	Examples
X	stimulus	message, picture, result of actions, etc.
γ	attribute of stimulus	medium, features of text and images, etc.
α, θ	parameters	attack and target parameters
α^p, θ^p	personal	age, gender, education, trust propensity, etc.
α^w, θ^w	work	years of service (YoS), role, domain, tasks, etc.
α^s, θ^s	setting	relevant goals, concurrent events, event time, etc.
Y	behavior	any action as a response to stimuli and parameters click, reply, download, etc.
t_{nth}	stage	current stage of the cognitive process

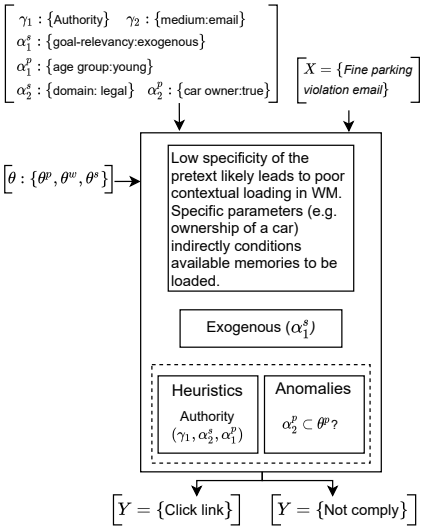
Parking fine phishing attack [256]

The first example is derived from a study on phishing susceptibility [256] and represents a simple phishing attempt whose pretext is a parking fine issued by (allegedly) a local police authority. The targets are nudged to click on a link in the phishing email; if this happens, the attack is considered successful. A representation of the attack using our framework is given in Fig. 3.2a and the verbatim text of the phishing email is provided in Appendix A.1 (Listing A.1).

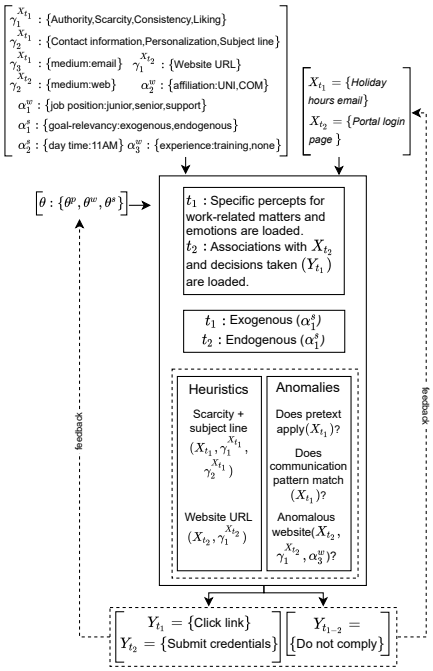
Stimulus & Parameters: The email is the stimulus X triggering the cognitive process of the target. Oliveira et al. [256] explicitly implemented an *authority persuasion technique* in the phishing email (modeled as the attribute of the stimulus γ_1 in Fig. 3.2a) and considered the attack parameters *age* (α_1^p) and *life domain* (α_2^s) in their experiment design. However, other parameters, such income, attention state, car ownership, etc., may also be relevant to the cognitive process of the victim. For simplicity, we include here attack parameters α_1^s=*goal-relevancy:exogenous* (as the stimulus is likely unrelated to the focus of the target when receiving it), and α_2^p=*car owner:true* (as the attacker assumes the target owns a car). These considerations emerge naturally from the attack description and pretext respectively given in [256].

Perception: At this point, the target's cognitive process automatically accesses past experiences related to the stimulus (e.g., dealing with bureaucracy, money concerns, previous decisions in similar contexts and associated emotions). The low specificity of the pretext is likely to cause only few or vague perceptive associations in the target. This means that the stimulus is likely to be only loosely linked to pre-existing memories due to the a-specificity of the message.

Attention: As in the attack simulation run in [256] the subjects do not expect to receive the provided stimulus, the pretext is unlikely to be linked to the current activity of the targets. Therefore, in most instances of the attack the attention block will process the stimulus as 'exogenous' to the current setting, matching the attacker expectation defined in α_1^s. Therefore, the initial elaboration is likely to be influenced by the use of less resource-demanding



(a) Un-targeted fine-parking phishing experiment [256]



(b) Tailored phishing experiment [61]

Figure 3.2: Examples of SE attacks from literature.

heuristics.

Elaboration: The attack implements the persuasion technique 'authority' (γ_1), exploiting the associated cognitive bias to increase the chances the target will comply with the email [73]. As shown in [256], authority is particularly effective when related to the legal domain and against young people, which are represented in our framework by attacker parameters α_2^s and α_1^p. Matching these parameters to the actual subjects receiving the stimulus will increase the chances the target will employ heuristic processing once directed here from the *Attention* block. On the other hand, the elaboration may occur with a higher amount of resources if an anomaly is detected. For example, if the subject does not own a car, i.e., there does not exist a target parameter matching α_2^p, an *Anomaly* is likely to engage more WM during elaboration (akin to re-reading a sentence that does not make sense at a first glance). Additional anomalies may be caused by the detection of a 'suspicious' URL in the email (e.g., as influenced by a subject's technical knowledge, a possible parameter in θ^p), or of an unknown sender.

Behavior: The attack succeeds if the target clicks the provided link, as per study design. In a real-world scenario, a new stage may be necessary to complete the attack (e.g., a phishing web page where to insert user credentials).

Discussion: This attack is rather unsophisticated as it relies on the fortuitous matching between attack and target parameters. The cognitive processing described in the *Elaboration*

step points out that mismatches between the parameters and the pretext may cause anomalies in the system that move the execution to the more cognitive intensive processing, which will lead to the failure of the attack. We note that the framework structure forces the identification of parameters (e.g., for attention and anomalies) that are not explicitly included in the original experiment design. This suggests that our conceptualization may be useful to identify factors (and limitations) in an experiment design; for instance, α_2^p can be a confounding variable in the attack.

Tailored phishing against organizations [61]

The second example concerns a study on tailored phishing against a university and a consultancy company where a bogus organization department asks employees to update their holiday schedule (this example refers to the work of Chapter 5). The pretexts are carefully designed to mimic internal communication patterns and cognitive exploits are employed to enhance the efficacy of the attack [61]. The targets are nudged to click on a link and enter their credentials on the fake company page; only submissions to the fake portal are considered as successful, making this a two stage attack. To capture this, we represent the stage in which a stimulus is used and write X_{t_i} to denote the stimulus used in the i-th stage of the attack. A representation of the attack using our framework is given in Fig. 3.2b and the verbatim text of the phishing email is provided in Appendix A.2 (Listing A.2).

Stage 1

Stimulus & Parameters: The email is the first stimulus triggering the cognitive process of the target (X_{t_1} in Fig. 3.2b). [61] tests four persuasion techniques: Authority, Scarcity, Consistency and Liking, represented as an attribute of the stimulus ($\gamma_1^{X_{t_1}}$). Additionally, every persuasion technique is enhanced with three notification methods: extended Contact information, Personalization towards the target and extended Subject line ($\gamma_2^{X_{t_1}}$ in Figure 3.2b).

The considered attack parameters are *job position* (α_1^w: junior, senior and support staff) and *affiliation* (α_2^w: university and company) which are the control variables as per experiment design. Other parameters, such attention state, work load, time of the day etc., may also be relevant to the process. We include here attack parameters α_1^s=*goal-relevancy:exogenous* (as the stimulus in the first stage, at t_1, is likely unrelated to the focus of the target when receiving it), and α_2^s=*daytime:11AM* (as the attacker assumes to hit a larger audience during the beginning of a work day). These considerations emerge from the experiment design and description given in [61].

Perception: In this stage the target's cognitive process automatically accesses past experiences related to the stimulus (e.g., dealing with organization matters, previous decisions in similar contexts and associated emotions). The rather high specificity of the pretext is likely to link perceptive associations in the target to relevant processes the subject is used to within the organization.

Attention: No assumptions on the subjects expecting or not expecting to deal with updating their holidays schedule in that time frame are provided in [61]. However, the pretext is unlikely to be linked to the current activity of the targets as scheduling holidays is a sporadic activity. Therefore, in most instances of the attack, the attention block will process the stimulus as 'exogenous' to the current setting. Hence, a low amount of cognitive resources is likely to be allocated to the initial elaboration of the stimulus.

Elaboration: This attack implements various persuasion techniques ($\gamma_1^{X_{t_1}}$), which exploit the associated cognitive biases to push the target to complete the decision-making process heuristically [73]. Following the experiment design in [61], the effect of these heuristics is

enhanced by their placement in the email (e.g., subject line, contact information, or signature), $\gamma_2^{X_{t_1}}$. On the other hand, the elaboration may occur with a higher amount of cognitive resources if an anomaly is detected, for example, when the subject has already completed a vacation schedule or the pretext does not apply to the target at all. For instance, as reported in [61], interns were 'immune' to the pretext due to their temporary position. Additional anomalies may be caused by inconsistencies with the usual communication patterns at the organization (e.g., as influenced by the subject's experience, e.g. θ^w: senior).

Behavior: The first stage succeeds if the target clicks the provided link and the second stage begins (t_2) with the phishing web page being displayed. Percepts and decisions made within this stage are retained and influence the target parameters and processing of the next stage (t_2).

Stage 2

Stimuli & Parameters: The second-stage begins with the website (X_{t_2}) led to by the link in the email. We consider the page URL ($\gamma_1^{X_{t_2}}$) as a relevant attribute for the processing the this new stimulus, since the attackers actively masqueraded the URL to look like legitimate [61]. The security knowledge of an employee can be represented with α_3^w to represent whether training has been administered.

Perception & Attention: Similarly to $X_1^{t_1}$, context is maintained with additional percepts concerning the displayed web page, like page contents and layout. We assume the attention deployed in this stage to be endogenous α_1^s because the subject may be actively engaged with the stimulus and have a defined goal in the `WM` at this point of the attack.

Elaboration: Although endogenous control is exerted, the exact replica of the page layout and design should accommodate the heuristic processing as the user may be habituated to login to the organization's portal [357]. However, an anomaly can be generated by processing the URL bar of the browser, or due to discrepancies with any other relevant previous percept or memory regarding the present stimulus (e.g., page formatting, 'lock' in the url bar, etc.). On the other hand, these effects also depend on previous knowledge of the user regarding general Internet security practices, possibly as influenced by received training (α_3^w).

Behavior: At this stage, the attack is successful if credentials are submitted in the bogus web portal.

Discussion: Unlike the previous case, the examined SE attack implements a tailored context for the targets in terms of a higher amount of conceivably matching work parameters. The investigators employ a set of persuasion techniques and delivery methods to favour heuristics-driven elaboration and increase the odds of success. The framework's explicit representation of anomalies allows to reason on the effects of a highly specific pretext on targets' cognitive processes: the α_2^w parameter (university vs. company) leads to different outcomes with respect to α_1^w (junior vs. senior vs. support) where the lack of knowledge of internal organization processes makes junior employees less capable in identifying anomalies in the communication [61]. Further, the two-stage break down of the simulated attack can make it easier to

isolate factors that may not have been considered during the experiment design: what is the influence of a mimicked URL versus a random one with the chosen pretext and parameters?

3.3.2. Breakdown of two real SE attacks

NGO spear-phishing [47]

This example is a tailored spear phishing attack against an NGO, where the email replays an actual announcement about a conference in Geneva and was edited by the attacker to indicate that all fees would be covered by the organizers and encourages to open an attachment [47]. A representation of the attack using our framework is given in Fig. 3.3a and the verbatim text of the phishing email is provided in Appendix A.3 (Listing A.3).

Stimulus & Parameter: The email pretext clearly revolves around the human rights topic in the context of NGOs and is anchored to two specific themes regarding the Uyghur population and a conference in Geneva. In particular, the attacker assumes the subject's work parameters (α^w) to map the professional context at a given NGO, and the setting parameters to represent the assumed attentional state (α_1^s), the conference date (α_2^s) and the time of the day when the email is sent, to reflect working hours (α_3^s). The stimulus' attributes comprise the impersonation of the sender (γ_1) and an invitation with covered costs (γ_2) (a trigger for the reciprocity persuasion technique commonly used in advertisement [73]).

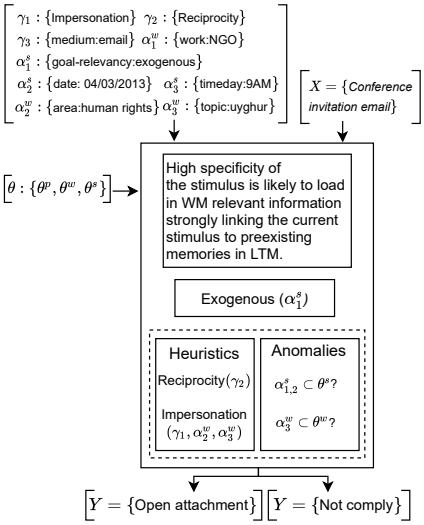
Perception: Given the high specificity of the stimulus, perception will yield the loading of a rich context in subjects that match the parameters. A rich set of percepts and cached computations provides the attacker with a wider attack surface, e.g., to trigger biases and to exploit heuristics related to that context.

Attention: We have no information on whether the targets are focused on actions related to that specific stimulus when it is processed; from the discussion provided in [47], we assume exogenous attention for most targets, which fosters the use of *Heuristics* later during elaboration in Fig. 3.3a.

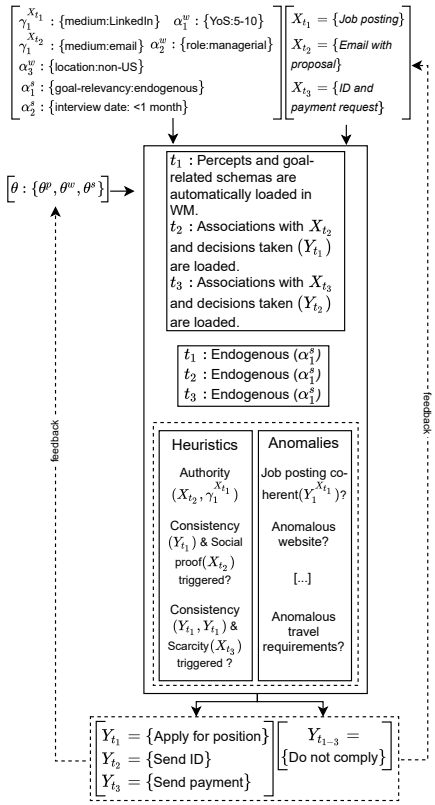
Elaboration: Three evident features of the loaded information aim to exploit the *Heuristic* processing: *(i)* citing different organizations and topics that the target presumably encounters frequently (α_1^w, α_3^w) exploits the availability heuristic; *(ii)* the invitation itself (α_2^s) and the covered travel costs (γ_2) appeal to the reciprocity heuristic; and *(iii)* the detailed contact information provided in the email boosts the validity of the messenger as an authoritative source (γ_1, α_2^w, α_3^w). These and other attack parameters (e.g., α_3^s) aim to facilitate as much as possible the reliance on *Heuristics* and compete against any other cue that might cause an anomaly or mismatch, like an inaccurate conference date (α_2^s) or anomalous timing for a work-related email from that source (α_3^s).

Behavior: The attack succeeds if the target decides to open the attachment as it contains an exploit leading to malware execution.

Discussion: A critical feature of this SE attack is the specificity of the pretext in relation to the experience of the targets. For example, a match of attack parameter α_2^s (i.e., whether the



(a) NGO spear-phishing attack [47]



(b) LinkedIn multi-stage attack [15]

Figure 3.3: Examples of SE attacks from real-word scenarios.

target has registered to the conference) with the actual experience of the target would likely positively reinforce the heuristic judgment. Importantly, were these parameters wrongly calibrated by the attacker, an anomaly would be likely triggered, causing more resources to employed for processing, thus thwarting the attack altogether. The attack flow shows that, unlike the first example (Sec. 3.3.1), a tailored pretext requires a large set of baseline parameters aligned with the target's context to enable the attack in the first place, similarly to what is discussed in [132]. When this necessary requirement is achieved, the attacker can further develop the attack (e.g., including cognitive exploits like in the second example, Sec. 3.3.1), to keep the victim's processing anchored to *Heuristics*. Whereas a large attack parameter space increases chances of success when well calibrated, the framework suggests that this also increases chances of mismatch, which may backfire and lead to attack failure. With this representation at hand, our framework can potentially enable the design of an ordinal metric to sort similar attacks in terms of matching parameters, amount of knowledge on the targets, akin to [281], and usage of cognitive exploits, e.g. [347].

LinkedIn multi-stage attack [15]

The last example is the case of a highly-targeted spear-phishing campaign against non-US white collar workers on LinkedIn [15], who are offered an appealing job position in the US. Prospect candidates applying for the job are first asked to provide documents and personal details (including a copy of their passport for VISA reasons), and then a payment for the anticipated (fake) traveling costs. Fig. 3.3b shows the application of the model and the attacker messages are provided in Appendix A.4 (Fig. A.1 and Listing A.4). This attack evolves through three stages (t_{1-3}), in which different messages are exchanged with the target.

Stage 1

Stimulus & Parameters: The initial stimulus X_{t_1} is the job offer post the subject is actively engaged with, i.e., the task is in her goal stack (represented by attack parameter α_1^s). The stimulus is tailored for a precise set of subjects, that is, experienced managerial workers not located in the US (represented by attack parameters α_{1-3}^w). The communication medium (i.e., LinkedIn) is represented by the stimulus attribute $\gamma_1^{X_{t_1}}$.

Perception & Attention: In the perception step, a specific and rich context is retrieved and readied to processing. Since the percepts and loaded associations are assumed to be goal-related, α_1^s, endogenous attention is likely triggered. Therefore, the elaboration will likely make use of more cognitive resources.

Elaboration: While using more WM, the target's cognitive resources are focused on the job post, this engagement may last only for a limited time period: the stimulus is delivered on the LinkedIn platform, a trusted source for job postings, and the job description is well curated and points to an existing website matching the LinkedIn company profile of the company advertising the job posting. These attributes of the stimulus all act in unison as 'heuristics' for legitimacy, pushing execution towards heuristic processing. Further, the job description advertises attractive job conditions and benefits, including insurance, leave periods to visit family abroad (after moving to the US), and a company car, further reinforcing biases.

Behavior: This stage of the attack succeeds if the target decides to apply for the job position. The decision and associated judgments made with the resource intensive WM (Y_{t_1}) are automatically stored in LTM, and will be made available for later use.

Stage 2

Stimulus & Parameters: In the second stage of the attack (t_2), the applicant is contacted via email (Listing A.4) with high promises (confirmation of eligibility, highlighting the importance of the role and explanation of the benefits) and low perceived costs (request to provide IDs for VISA/application). In this stage, the communication medium is an email, represented by stimulus attribute $\gamma_1^{X_{t_2}}$.

Perception & Attention: When stimulus X_{t_2} is processed in *Perception* (at t_2 in Fig. 3.3b), previous experience, decisions and associated judgments (Y_{t_1}) are recalled from LTM. These are key aspects to foster deception in this and later stages of the attack as they produce reinforced schemas based on previous experiences that the target will rely on to form upcoming judgments and decisions. The stimulus is still goal-related and, thus, endogenous attention is allocated, leading to a higher usage of WM.

Elaboration: The loading of the percepts in the previous step enables a set of heuristics related to the previous, implicit commitment made in the first attack stage. This is well aligned with the attacker's objective to keep the processing as much as possible towards using *Heuristics*, where X_{t_2} can exploit a number of cognitive biases. The foremost bias exploited in X_{t_2} pushes the subject to remain consistent with their previous decisions (Y_{t_1}) [73]. Additionally, Social proof, Scarcity and Authority can be exploited by the attacker, with the latter two supported by attack parameter α_2^s (i.e., little time ahead of the interview), and stimulus attribute $\gamma_1^{X_{t_1}}$ (i.e., a trustworthy source) respectively. At this point, unless an anomaly is triggered, the heuristic processing reaches a decision whether to continue with the application.

Behavior: This stage of the attack succeeds if the target decides to send the required documents (note that the attacker can already extract value from the attack in the form of ID theft from the passport scans and submitted subject details). As in stage 1, the decision and associated judgments of this stage (Y_{t_2}) are also stored in LTM for later use.

Stage 3

With the increasing strength of percepts characterizing target's previous commitments (Y_{t_1} and Y_{t_2}), the attack enters in its third and final stage (t_3) in which a payment is requested (cf. bottom of Listing A.4). Stimulus X_{t_3} is processed as in the previous stages, now with added support to *Heuristics* in form of Consistency (with y_{t_2}) and Scarcity biases. The latter is achieved by setting the date of the alleged interview relatively close to when the communication happened (α_2^s) – and with the requirement of getting a VISA in time despite the upcoming Christmas vacations. At this point, the most relevant anomaly that may jeopardize the decision to comply are the travel constraints (the requirement to book through the affiliate travel company). If the commitment to undertake a positive decision overcomes the costs of compliance [15], the target will most likely comply with the attacker's request and send the payment (Y_{t_3}).

Discussion: We showed how the model allows one to consistently break down complex attacks into essential steps characterizing the target's cognitive processing of the attack. By highlighting the interaction of multiple stages, we can study the effects of the attacker's strategy, such as the trade-off between target commitment and (escalating) attacker requests (e.g., to define when best to advance a payment request as opposed to asking for additional personal details). We also observe the tactics used to elicit new information, such as applying the Social proof persuasion technique in the second stage to gain a stronger 'foot hold' on the target's side (i.e., it is usual business for large companies to arrange travel and ask for documents). It is worth noting that these considerations are in line with the art of deception whose aim is to reduce suspicion in the target's mind [233, 321]. Importantly, studying the means by which a target's processing flow can be deviated from the processing 'desired' by the attacker may open the way to new training techniques or decision support systems, for example actively detecting the 'escalating' nature of complex SE attacks as revealed by the proposed framework.

3.4. Discussion

3.4.1. Implications for research

We expect the framework to be used by researchers to systematically identify shortcomings of simulated attacks and experiments, such as isolating factors that are difficult to recognize without a reference to the features of human cognition (e.g., spotting anomalies), and when constructing pretexts and keeping track of targets' context (e.g., the matching of parameters). The framework can be applied to studies that simulate SE attacks to map the effects 'modeled' in the experiment design (i.e., what the study aims to investigate) on the components of the cognitive framework. The framework might aid research over several dimensions:

The parameter space, to assure the modelling of a realistic attacker that can match or measure the attack and target parameters, as well as factors concerning the context of their targets that may influence the outcomes (e.g., exogenous or endogenous attention due to subject variables). Similarly, pre-existent memories and experiences may be employed in empirical settings to evaluate the effects of *percepts* (in the 'perception' block) on the unfolding decision-making.

Stimulus engineering over pretext and attributes. The engineering of an artifact goes beyond the mere presence of triggers for cognitive biases, and considers additional features such as the effect of the message medium on perception. For example, emails are often associated to phishing, while LinkedIn posts may not. What are the expected interactions between the stimulus attributes, and the characteristics of the receiver (subject parameters)?

Attack execution and effect measurement. The framework might help in identifying the key modulating aspects impacting the execution of the attack, for example what type of central attention is expected in the subject when delivering the artifact. The iteration and modifications of attack/subject parameters in multi-stage attacks can also be 'modelled' following the proposed framework: whereas no formal empirical work has been carried out to date on this aspect, the conceptualization proposed by our model shows that, as in the LinkedIn attack example, multiple target-attacker interactions can significantly modify the parame-

ter space across attack stages. The framework might further help researchers in structuring post-experiment measurements, for example by means of surveys, to assess the effects of the attack/stimulus at the different levels of a target's cognitive processes. These include the usage of heuristics, detection of anomalies, but also the possible presence of percepts and memories that affect the computation.

Designing and assessing defensive policies and training. Training activities can be aimed at different levels of a cognition process. For example, from identifying known biases to increasing the chances of an anomaly triggering. Defensive policies might further benefit from this conceptualization by investigating if and how the subject or organizational parameter space can be tuned to increase the chances for anomalies to occur. For example, by introducing specific target parameters (e.g., language) in everyday communication patterns inside the organization that are not easily matched by outsiders, or that is incompatible with the triggering of innate biases (e.g., *authority*).

3.4.2. Implications for practice

The application of the framework to sophisticated attacks against NGOs and LinkedIn users revealed that the analysis of complex attacks can be simplified and structured to analyze and compare different attacks, their techniques, and execution conditions. We expect that breaking down sophisticated attacks can help to get insights on the causes behind their effectiveness and to devise new detection methods. For example, during the analysis of security incidents, by including potential target-related contextual variables affecting their decisions and behavior in an attack. From a threat intelligence point of view, the forced identification of parameters of an attack, and the match thereof, might help devise risk metrics for different typologies of attacks, for example, based on their level of sophistication. Identifying parameters of an attack may be especially relevant when dealing with internal threats, like ex-employees or undetected compromised accounts, since insider information can be exploited to carry out effective attacks [7]. Importantly, the breakdown of real attacks is expected to facilitate researchers and practitioners alike to keep track of innovative or previously unseen attack techniques, contextualizing and isolating those in the overall cognitive process, opening the way to better training, policies, and research targeted at measuring related effects. Finally, our framework can be useful when designing security systems to reduce the opportunity for cognitive shortcomings to trigger undesired behaviour, similarly to procedures used in design and human computer interaction [84].

3.5. Related Work

Systematic attempts to explore and characterize SE attacks through the lenses of cognition are so far unstructured and limited in number. Our goal requires a framework able to capture attack characteristics and cognitive features at a sufficiently high level of abstraction in order to be applicable to virtually any SE attack. At the same time, the scope of such a framework needs to be neither too narrow nor too broad to include the relevant features of a SE attack while not being so general as to be unspecific to the relevant SE constructs. To the best of our knowledge, only a few works provide a framework related to human cognition which is applicable to study and analyze SE attacks, albeit at various levels and with different scopes:

Cranor [84], Dolan, Hallsworth, Halpern, King, and Vlaev [98], Montañez, Golob, and Xu [238], Tetri and Vuorinen [332]. Table A.1 in Appendix A presents a detailed comparison between the features covered in this framework and the other identified frameworks.

Montañez, Golob, and Xu [238] map a selection of short- and long-term human factors affecting a subject's susceptibility to SE attacks, such as workload, experience, knowledge, and culture, on a basic and high-level framework of human cognitive functions, that is, memory, perception, attention, and decision-making (including heuristics). Cranor [84] proposes a framework for reasoning on the root cause of security failures attributed to human error in the loop of a (computer) system, based on the Communication-Human Information Processing (C-HIP) model from the warning science literature [83]. The MINDSPACE framework proposed by Dolan, Hallsworth, Halpern, King, and Vlaev [98] is a framework for public policy-making that summarizes the most robust influences for behavior change across a variety of contexts. Elements of the MINDSPACE framework that concern how behavior can be influenced include message characteristics, contextual variables (e.g., the messenger), and features of automatic processes of judgment, such as priming, salience, and heuristics. Tetri and Vourinen [332] introduce a conceptual framework for SE that aims to extend the study of persuasion in SE attacks to include additional parameters of the targets and environment, such as organizational settings and security policies.

The framework presented in Montañez, Golob, and Xu [238] is the closest to the framework described in this chapter, as both are grounded on existing theories of cognition and share a similar structure and mechanisms of cognition. The scope of the framework in [238] concerns a high-level application of principles from the cyber-security domain on the cognitive psychology field, such as the influence of cognitive factors, memory, and attacker effort on behavior. However, due to the lack of an explicit mapping to SE attacks, this framework would leave room for interpretation on how to instrument it to carry out an analysis of various SE attacks and, therefore, unsuitable for our purposes. In contrast, the current framework has been specifically proposed to analyze real and simulated attacks from the literature and evaluate research results by decomposing and comparing different SE attacks in terms of cognitive features. Similarly to our framework, Cranor [84] describes the processing of information by a human receiver whose behavior is dependent on a set of processing steps and mechanisms, personal and environmental variables. However, the scope of their work is to facilitate the design of secure systems that rely on humans, such as an anti-phishing tool providing warnings, with an explicit focus on comprehension and retention of knowledge. Instead, our framework aims to contextualize SE attacks from the point of view of cognition where the received inputs and processing steps are related to the attack itself. For example, knowledge retention and transfer are key mechanisms in [84] whereby a human receiver that processes warnings, such as of an anti-phishing tool, applies this knowledge to future warnings; however, they are not central to the processing of SE attacks. The application of our framework to real SE attacks shows that this framework is able to capture the main aspects concerning knowledge retention and transfer that are relevant to the study of SE attacks. The remaining frameworks related to human behavior and/or cognition aspects [98, 332] either do not relate those aspects to SE attacks [98], or focus on a narrow set of cognitive features, such as parameters or persuasion techniques [332]. By contrast, attack characteristics and cognitive processes are the key elements of the framework proposed hereby.

In Table 3.2, we show that the features considered here are necessary to characterize and analyze SE attacks. This is supported by our analysis of the identified frameworks related to human cognition that can be applied to study SE attacks and, in general, to the cyber-security domain. Indeed, as shown in Table A.1 in Appendix A, all cognitive features relevant to SE attacks have been covered by at least one identified framework other than the current. Yet, no framework (except ours) covers all the relevant features, making them less suitable to evaluate and contextualize research results. Overall, the levels of abstraction in the different works vary to a point where some are too low-level to represent any SE attack, and other have either a too narrow scope to capture relevant SE features, or only partially overlap with the domain of social engineering. For example, the framework in [238] can virtually accommodate any SE attack, but it lacks mechanisms to sufficiently characterize them; the framework in [84], on the other hand, describes a rich set of processing steps and factors, but the scope only partially overlaps with SE; the framework in [98] is too broad and barely overlaps with SE, and the framework in [332] focuses on a narrow set of SE characteristics. This makes it difficult to use any of the discussed frameworks to analyze SE attacks in the context of human cognition.

3.6. Conclusion

In this chapter, we presented a cognitive framework to dissect and characterize SE attacks. The framework is based on well-established theories and models of human cognition drawn from the field of cognitive sciences, such as the Working Memory model and Dual-process theories. The relevant cognitive processes are mapped to the SE domain allowing us to relate effects and attack techniques to specific cognitive features and processes of the targets, for example, linking perception or attention to the conditions of an attack. The framing of attack parameters (i.e., attacker assumptions on the target) and target parameters (i.e, characteristics defining the target and their context), and the match thereof, provide a structure to reason on the tailoring degree of the attacks and the related effects, such as the influence of pretext on the target's decision making. We showcased the proposed framework against four attacks (from realistic to real, and from general to highly-targeted), illustrating its application both for experimental design (or empirical SE research), and as an instrument to characterize SE attacks in the wild.

As outlined in the Chapter 1, the lack of a structured understanding of socio-technical aspects of the SE domain makes it difficult to interpret and compare experimental results. The work presented in this chapter, therefore, serves as an instrument for the characterization of the SE attack surface, and to evaluate and contextualize research results. This provides a contribution toward answering RQI. In the next chapter (Chapter 4) we employ the framework to answer RQII, enabling us to identify gaps and open research questions in empirical SE research.

4

The vastness of SE attack surface: gaps between human cognition and empirical research

The interdisciplinarity of the SE domain creates crucial challenges for the development and advancement of empirical SE research, making it particularly difficult to identify the space of open research questions that can be addressed empirically. This space encompasses questions on attack conditions, employed experimental methods, and interactions with underlying cognitive aspects. As a consequence, much potential in the breadth of existing empirical SE research and in its mapping to the actual cognitive processes it aims to measure is left untapped. By means of the framework presented in Chapter 3, we carry out a systematic review of 169 articles investigating overall 735 hypotheses in the field of empirical SE research, focusing on experimental characteristics and core cognitive features from both attacker and target perspectives (see RQII). Our study reveals a series of open gaps in relation to how SE attacks are reproduced in experiments and the coverage of the exploitable attack surface by the current body of research, among others. Our findings on current SE research dynamics provide insights into methodological shortcomings and help identify supplementary techniques that can open promising future research directions.

This chapter is originally published as P. Burda, L. Allodi, and N. Zannone, "Cognition in Social Engineering Empirical Research: a Systematic Literature Review", In *Transactions on Computer-Human Interaction* (ToCHI), ACM, 2023

4.1. Introduction

Humans are a critical component of any computer system and, as such, are part of a system's attack surface. The 'human vulnerabilities' exploited by SE attacks are ingrained in human cognition and, as they are shared across all 'human targets', these vulnerabilities represent a rather stable attack surface, allowing attackers to avoid the complexity and costs associated with deploying malware-based attacks [14, 147, 238]. Whereas a number of surveys have already extensively analyzed defense methods against SE attacks [121, 147, 286, 303], a comprehensive overview of studies analyzing the SE attack surface is still missing. Critically, a clear empirical understanding of the overall attack surface is necessary to design and test defensive techniques effective against emerging SE attacks.

In Chapters 2 and 3, we observed an increasing sophistication of the techniques adopted by attackers to exploit human-based vulnerabilities, moving away from simplistic phishing campaigns targeting the 'mass' of Internet users, into tailored multi-step attacks leveraging target weaknesses and target-specific information. [15, 39]. These attacks are often tailored against specific organizations or groups of people, exploiting the specificity and characteristics of their targets [47, 148]. Therefore, research in this area needs to capture multiple perspectives from a variety of disciplines, such as cognitive and social psychology, to grasp the nuances of interactions between the technical aspects of an attack and the cognitive dimensions characterizing its human element. However, as seen in the previous chapters, the interdisciplinarity of the SE domain makes it particularly difficult to identify gaps and open research questions as well as to interpret experimental results [15, 147, 332, 368]. Efforts to study the cognitive aspects related to the SE domain are so far relatively unstructured, which hinders a coherent interpretation of cognitive effects, replication of experiments, and evaluation of gaps.

In this chapter, we present a systematic review of 169 research articles in the field of empirical SE with the aim of identifying and characterizing open gaps between the features of human cognitive processes and empirical research in SE. We employ snowball sampling on an initial collection of relevant literature obtained from the Scopus database and employ the cognitive framework of Chapter 3 to derive the foundational cognitive dimensions evaluated by the extant literature. Our criteria cover the experiment setup, the characteristics of the simulated SE attack, the target's cognitive processes and characteristics, and the interactions between such variables.

Our study shows that most experiments only partially reflect the complexity of real SE attacks and investigate only a small portion of the overall attack space (e.g., *single-step-mono-modal* attacks, as opposed to more sophisticated – and increasingly more relevant [15] – *multi-step-multi-modal* attacks). Moreover, our review reveals that the exploitable SE attack surface appears much larger than the coverage provided by the current body of research. For example, despite their high relevance for both attack design and defense, factors such as targets' context and cognitive processes are often ignored or not explicitly considered in experimental designs. Similarly, the effects of different pretexts and varied targetization levels (i.e., to what extent an attack was adapted to the recipient) are overall marginally considered. We find that the literature is overall focused only on a few experimental setups, it lacks a common reference for attack targetization and the experimental outcomes are rather inconsistent in

defining when a SE attack is deemed successful. These issues limit the explanatory power of results, the reproducibility of experiments, and the innovation of experiment designs. Based on our findings, we report promising, interdisciplinary future research directions, as well as still-untapped resources for the design of innovative experiments and effective defensive mechanisms.

The chapter is structured as follows: Section 4.2 introduces a brief background on empirical approaches adopted in SE research relevant to the analysis. Section 4.3 describes our methodology for data collection, lays out the research sub-questions, and derives the criteria used in the analysis. Section 4.4 presents the results of the analysis and Section 4.5 discusses our findings.

4.2. Background and Related work

4.2.1. Empirical approaches adopted by SE studies

The study of cognitive processes in the SE context is typically carried out through experiments that aim to reproduce real attacks and measure the emergent behavior of the involved participants. A participant's behavior is the result of their cognitive process influenced by the attack stimuli as well as their cognitive and contextual characteristics. Therefore, these constitute essential factors to be measured, controlled, or evaluated in the experiment.

Different types of experiments have been conducted to study the effects of SE attacks, ranging from *field* and *laboratory experiments* to *observational studies*. The choice of the experiment type usually depends on the scope of the study and availability of resources: if environmental variables are of interest, a field experiment may be more apt to study their effects on behavior, whereas a laboratory experiment allows researchers to isolate variables that are too difficult or impossible to control otherwise. Field experiments, such as unannounced embedded phishing training [193], are carried out within the natural environment of the participants, to retain ecological validity. Laboratory experiments are, by contrast, used to measure the effects of specific contextual factors, such as user interfaces [322], or factors related to cognitive processes, such as participants' gaze [234]. However, the outcome of laboratory experiments may not be easily generalizable to real-world settings, notably due to ecological constraints. Observational studies involve the (retrospective) observation of the effects of risk factors or treatments, such as in case-control or cohort studies [16], where any independent variable is out of the control of the investigators. Other types of experiments commonly adopted in empirical SE research are *surveys* and *interviews*, either by themselves, e.g. [96], or complementary to lab and field experiments [378]. These study types are particularly relevant when an effect cannot be observed directly, for example, the rationale behind the detection of an anomaly [378] or decision making [360].

The process to set up an experiment largely matches that of real attacks (cf. Section 2.1 and Fig. 2.1). In the *preparation phase*, researchers define the experiment objectives and the hypotheses to be tested and, based on them, determine *control variables* (i.e., independent variables used to control for confounding effects), *treatments* (i.e., independent variables that are manipulated by the investigator), and *outcome variables* (i.e., dependent variables that may be impacted by the independent variables). This usually involves identifying the rele-

vant attack parameters along with the type of stimuli and their attributes to be used in the experiment as well as determining the behavior of the studied targets to be measured. The preparation phase also aims to identify the attack environment and potential victims. Thus, this phase also emulates the reconnaissance phase of a real attack, in which the attacker identifies targets of interest, as shown in Fig. 2.1. The *subject selection* phase encompasses the selection of the actual targets of the experiment, e.g., students of a university or employees of a company, based on the identified attack parameters.

The *artifact construction* phase concerns the realization of the stimuli (and often involves the deployment of the infrastructure) used in the experiments based on the attack parameters and hypotheses to be tested. This might include, for instance, the implementation and deployment of a phishing website where the targets should submit their credentials, thus determining how the targets' behavior is recorded (orchestration phase in Fig. 2.1). The stimuli have to be constructed in such a way they reflect the objectives of the study and the modeled threat. To this end, the artifacts may be adapted to the subjects and include cognitive exploits or other features. The *adaptation* of the artifacts to the targets is particularly critical for the final outcome of an attack: experiments and real-world cases indicate that attackers can leverage the information on targets to build tailored messages, achieving high success rates both in absolute terms [15, 62] and relative to those of non-tailored attacks [63, 120]. The *execution* phase concerns the delivery of the stimuli to the targets; in the *measurement* phase the experimenter measures the outcome of interest, typically in terms of participants' behavior (e.g., clicks on the link, credential submissions) or other indirect effects.

4.2.2. Related Work

Previous research in the field of SE has been summarized in several literature studies, whose focus ranges from an analysis of attack characteristics and victims' underlying cognitive processes to a review of the proposed defense techniques and of the performed experiment designs. Pfleeger and Caputo [276] survey behavioral science findings relevant to cybersecurity, which partially cover cognitive process features, for example, elaboration and behavior. Darwish et al. [85] investigate the relationship between victims' characteristics such as demographics and personality traits (parameters) and phishing attacks, along with an analysis of existing detection techniques. Heartfield and Loukas [147] propose a taxonomy of semantic SE attacks along with their characteristics and review defense techniques, similarly to Salahdine and Kaabouch [303] and Purkait [286]. Tetri and Vourinen [332] introduce a conceptual framework for SE to analyze attack characteristics, parameters of targets and setting, and the execution of SE attacks. Montañez et al. [238] map existing studies on various aspects of SE attacks into a basic and selective framework of human cognition functions and delimit their considerations to artifact construction, short and long-term cognitive factors, attention selection, and behavior. Sommestand and Karlzen [320] analyze phishing field experiments by looking at experimental variables, results (susceptibility rates), and experiment design features, such as explicit hypotheses, control variables, etc. Finally, Franz et al. [121] present a taxonomy of phishing interventions for usable security that comprehends the design of training experiments, including the type of training, attack vectors, and some contextual factors. The authors also review the user interaction problem that touches the relevant cognitive processes, such as perception and elaboration.

Table 4.1: Literature studies on Social engineering with their coverage (● means "covered", ◐ "partially covered", ○ "not covered").

		Pfleeger & Caputo [276]	Darwish et al. [85]	Heartfield & Louka [147]	Purkait [286]	Tetri & Vourinen [332]	Montañez et al. [238]	Sommestand & Karlzen [320]	Salahdine & Kaabouch [303]	Franz et al. [121]	This literature review
Attack char.	Attack vector	○	○	●	◐	◐	○	◐	●	●	●
	Attack targetization	○	○	●	○	●	◐	●	●	○	●
Cognitive processes	Parameters	◐	◐	○	○	●	●	●	○	◐	●
	Perception	◐	○	○	○	○	○	○	○	●	●
	Attention	◐	○	○	○	○	●	◐	○	◐	●
	Elaboration	●	○	◐	○	◐	●	○	○	●	●
	Behavior	●	○	○	○	○	●	●	○	◐	●
Defense tech.	Prevention	○	○	●	●	◐	○	○	●	●	○
	Detection	○	◐	●	●	○	○	○	●	●	○
	Mitigation	○	○	●	●	○	○	○	●	●	○
Experiment design	Experiment type	○	○	○	○	○	○	◐	○	●	●
	Preparation	○	○	○	○	◐	○	◐	○	◐	●
	Subject selection	○	○	○	○	○	○	●	○	○	●
	Artifact construction	○	○	○	○	◐	○	◐	○	○	●
	Measurement	○	○	○	○	○	○	◐	○	◐	◐

Table 4.1 presents a comparison between our literature review and other surveys. Only the survey in [320] relates attack characteristics and cognitive processes, although only implicitly and partially. By contrast, our analysis explicitly considers target and attacker perspectives and relates them to each other, but without systematically analyzing the effect size on SE susceptibility (we refer to Section 4.3.3 for the motivations underlying this choice). Other literature reviews [238, 276, 332] focus on human behavior and cognition aspects, but do not relate those to SE attacks nor examine aspects related to the experiment design. The survey in [121] relates parts of cognitive processes and experiment design characteristics to SE attacks, but focuses mainly on prevention and user interaction aspects. The other reported surveys (i.e., [85, 147, 286, 303]) do not treat cognitive-related aspects nor look into the experiment design.

4.3. Systematic Literature Review process

4.3.1. Research Questions

Our overarching goal is to advance the body of knowledge in the SE domain by identifying and characterizing open gaps between the features of human cognitive processes and empirical research in Social Engineering. We refine RQII in a number of specific research questions covering the empirical approaches adopted by SE studies (RQ1-4), and the studied cognitive effect (R5 and related subquestions, and RQ6). The relation between the RQs and the overall process of attack engineering and experiment design covered by this literature review is depicted in Fig. 4.1.

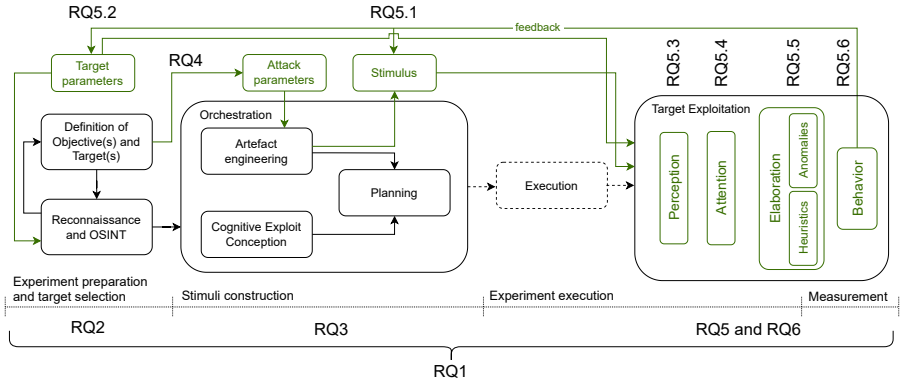


Figure 4.1: Overview of the research questions (cf. with Figure 2.1).

Empirical approaches adopted by SE studies

We first explore the various empirical methods adopted in the literature to understand how researchers reproduce the SE attack process described in Section 4.2 and the different aspects of real attacks that have (or have not) been covered with such methods. Although an experiment design can be very nuanced, our aim is to capture the context of a study in terms of used empirical methods, sampled population, artifacts, and their tailoring to the subject population (reflecting 'Attack characteristics' in Table 4.1). The following question is aimed to capture the empirical methods used in the SE literature:

RQ1 *What empirical methods have been adopted to study cognitive effects in the SE literature?*

A general limitation of experiments involving human subjects is the relation between the sampled population and the external validity of results [304]. In the SE context, this is particularly relevant because experiment outcomes are largely affected by the characteristics of the target population [61, 320]. The choice of the target population also reflects the subject selection and reconnaissance stages of an SE attack process (see Fig. 2.1). This is reflected in the following question:

RQ2 *Which subject populations have been considered to sample targets in empirical SE literature?*

Artifact engineering is a critical step of any SE attack (cf. Fig. 2.1); thus, the artifacts used in an experiment play a crucial role in the understanding and interpretation of its outcomes. For example, it has been shown that the success rate of SE attacks is largely influenced by the stimuli type and media used in the attack [150, 209]. We therefore ask:

RQ3 *What types of artifacts have been considered for the delivery of social engineering attacks in empirical SE literature?*

The tailoring of artifacts towards the targeted population (i.e., the alignment between attack and target parameters, in terms of the cognitive framework presented in Chapter 3) is becom-

ing an increasingly common practice in real-world attacks [147, 303] and has been shown to have a significant impact on the outcomes of SE experiments [132, 320]. The next question aims to explore the relationship between the target population and phishing artifacts to shed light on the study context and the level of attack targetization investigated in the literature:

RQ4 *To what extent are SE artifacts tailored to the experiment subjects in empirical SE literature?*

Cognitive features

Attackers are known to engineer their cognitive exploits, i.e., the construction of a believable identity and pretext, to increase the effectiveness of SE attacks [15]. Therefore, the empirical investigation of cognitive effects is a necessary step for understanding the underpinning mechanisms that lead to victimization.

The long-standing problem of why SE attacks work and the inability of current solutions to neutralize such attacks have spurred researchers from information systems, human-computer interaction, and computer security to explore and isolate human-related factors affecting subject deception [88, 99, 380]. In Chapter 3, we have seen how these represent the "SE attack surface" which encompasses the ways an attacker can deceive the target to accomplish her goals and how it can be characterized along the dimensions of: stimuli attributes, target characteristics, and the contextual situation around the target (*personal target parameters*, *work-related target parameters* and *setting-related target parameters*, and cognitive processes; refer to Table 3.2 for the description of each cognitive feature and Table 3.3 for examples of target parameters). Our aim is thus to understand which factors of the SE attack surface have been investigated in empirical studies to analyze SE attacks. This leads to the following research question:

RQ5 *Which cognitive features of SE attacks have been tested empirically in the SE literature, and in which experimental settings?*

This question can be further refined based on the features of the cognitive process presented in Chapter 3:

RQ5.1 *What stimuli attributes have been investigated?*
RQ5.2 *What target and contextual characteristics have been investigated?*
RQ5.3 *What effects on perception have been investigated?*
RQ5.4 *What effects on attention have been investigated?*
RQ5.5 *What effects on elaboration have been investigated?*
RQ5.6 *What types of behavior have been investigated?*

On the other hand, effects at the level of a specific component of cognition may vary (e.g., being reinforced or neutralized) by the engagement of other components. These interactions have been reported in previous studies, for example perception manipulation leads to mixed effects on behavior [174, 268] and elaboration [129]. These interactions are also affected by the level of tailoring of artifacts [125, 150]. We, therefore, posit the following research question:

RQ6 *What interactions between cognitive features have been studied in empirical SE literature?*

4.3.2. Paper collection

To cover the wide and interdisciplinary SE landscape, we chose the Scopus database as the initial data source. Scopus is a large multidisciplinary database covering published material in the humanities and sciences. Compared to other databases (e.g., Web of Science), Scopus is among the databases that index the highest numbers of unique articles in computer science [67] and provides wide coverage of venues (journals ad conference proceedings) [106], including most of the top tier venues on security (e.g., IEEE Security & Privacy, ACM CCS, USENIX Security) and on human-centric security (e.g., ACM CHI, SOUPS).[1] Additionally, Scopus includes only peer-reviewed studies and offers a set of tools that allow one to limit the search to titles, abstracts, and keywords, and to subject areas (e.g., computer science), thus providing an efficient means for the lookup of relevant studies while keeping a high recall.

To answer the research questions presented in the previous section, we collected previous literature by building the following search query, which relates to social engineering, empirical research, and cognition:

```
TITLE-ABS-KEY ((”social engineering” OR phishing OR scam*)
AND (empirical* OR experiment*) AND (cognit* OR psycholog* OR
behavi* OR persua* OR influenc*)) AND (LIMIT-TO(SUBJAREA, ”COMP”))
```

We derived the specified keywords from our research questions, and included keywords `phishing` and `scam` to cover papers where social engineering is not mentioned explicitly.[2] `psychology` is included because it is a concept closely related to behavior, while `persuasion` and `influence` techniques are the main purpose of SE attacks. Finally, `empirical` and `experiment` are related to RQ1-4. The search query was executed on the Scopus database as of August 2021 on title, abstract, and keywords, and limited to papers published until December 2020 in the Computer Science subject area.

To be included in the review, a paper must satisfy the criteria in Table 4.2. The third criterion derives directly from RQ5 and is meant to cover aspects related to human information processing and behavior. The fourth excludes works not considering SE attacks, e.g., studies only focusing on training without a simulated SE attack. An example of a study that satisfies such criteria is a phishing susceptibility study written in English (1st criterion) where simulated phishing emails are sent (2nd and 4th criteria) and relevant behavior measured (e.g., clicking links in emails, 3rd criterion) [239]. A study that does *not satisfy* the criteria is an evaluation of phishing detection algorithms (not satisfying the 3rd criterion) [348] or a survey measuring the self-reported victimization and connected factors (not satisfying the 4th criterion) [69].

We applied snowballing to the papers retrieved from Scopus that met the criteria in Table 4.2

[1] From the best of our knowledge, the only top computer security venue not indexed in the Scopus database is the Network and Distributed System Security (NDSS) Symposium.

[2] We did not observe any substantial differences when executing the query with more specific keywords such as `vishing`, `smishing`, etc.

Table 4.2: Inclusion criteria for the SE literature.

Inclusion (a paper must)
Be published in English
Be an empirical study
Use principles and techniques from cognitive sciences as independent or outcome variables
Describe an SE attack

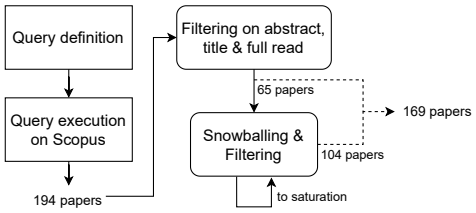


Figure 4.2: Paper selection process.

till saturation was reached, i.e., when the snowballing process yielded no additional papers.[3] This ensures our paper sample comprises relevant papers that may not be covered by the Scopus database and/or papers missed by the search query (e.g., due to the title or abstract non including the required keywords). The references gathered with the snowball procedure were looked up on Google Scholar and retrieved from the publisher's website. The summary of the whole procedure is shown in Fig. 4.2. The filtering step in the figure corresponds to the application of the criteria of Table 4.2 on the title and abstract first and on the full reading afterward. Similar articles by the same authors, such as conference papers extended into journal papers, were excluded and only the latest more extended versions were included in the literature review. Paper filtering and data extraction were performed by one of the authors and, in case of uncertain or ambiguous cases, the authors iteratively discussed and confronted until a consensus was reached. Out of the 194 papers initially found, 65 met the inclusion criteria for the analysis and after the snowballing saturation and filtering cycles additional 104 papers met the criteria, for a total of 169 papers included in the review.

4.3.3. Evaluation Approach

Empirical approaches

To characterize the study design (RQ1-4), we categorize the *study* types (i.e., field experiment, lab experiment, observational study, survey, interview), *population type* (e.g., general public, students, university staff, company employees), *attack vectors* or *stimuli* (e.g., email, voice, social network) and the degree of *attack targetization* studied in the selected literature. Similarly to [320], we classify attack targetization in: *individual* (I), *population* (P), and *generic* (G) to denote whether the attack employs information about a specific individual or a category of individuals. In addition, we use class (U) to denote that the provided information is insufficient to determine the degree of targetization. We determine the degree of targetization by analyzing the matching between the subject selection procedure (target parameter) and the subject parameters assumed in the attack (attacker parameter). Specifically, we considered the stimuli and treatments (the content, relevant attributes, such as the sender or pretext, and their variations) with respect to the intended recipients and the overall context described in the study. Lab experiments have often been carried out in online environments, such as crowd-sourcing platforms, and are broadly considered an extension of physical loca-

[3]When a paper references a paper stored in archives or in similar repositories, we checked whether the referenced paper has been published in a peer-reviewed outlet and considered the peer-reviewed version for our analysis.

tions [335]; therefore, we retain the label 'lab experiment' for such cases. In addition, many field and lab experiments include one or more surveys (and possibly interviews) as part of their data collection method. In these cases, we include the survey label in categorizing an article when such survey measures a distinct dependent variable that relates to the cognitive features, such as 'susceptibility awareness' [209] or 'reasons for behavior' [216], in contrast to, e.g., only demographics [119].

Cognitive Features

The criteria used to answer RQ5 and RQ6 directly stem from the features of the cognitive framework presented in Chapter 3. We apply the identified criteria to the experiment design of the selected studies and focus on tested hypotheses and/or research questions. We consider only quantitative or qualitative results that are *explicitly* reported in the results section of the paper. Specifically, for each hypothesis/research question in a paper, we identify *control variables*, *treatments*, and *outcome variables*, and map them to the features of our framework. As some variables (e.g., attributes of a stimulus) can be used to manipulate the cognitive process further down the cognitive pipeline, we also capture *indirect effects* whereby a manipulation can have cascading effects on other blocks (e.g., triggering a cognitive bias in Elaboration). This allows us to map the effects 'modeled' in the experiment design (i.e., what the study aims to investigate) on the components of the cognitive framework presented in Chapter 3. We do not capture the directionality of effects, as a direct (and fair) comparison is not possible, as it would require matching formulated hypotheses across different study designs (including subject groups, the domain of application, and artifact implementation); differently, in this study, we are interested in capturing the relation and mapping of these hypotheses to the relevant cognitive features.

Overall, we identified 792 hypotheses, of which 57 were removed because not relevant (articles fulfilling the selection criteria of Table 4.2 can contain hypotheses that are out of scope, e.g., measuring task duration, memory performance), for a total of 735 hypotheses included. It is worth noting that, when no hypothesis/research question was explicitly provided, we derived them from the experiment description and/or method section.

Stimuli attributes. We report the attributes describing stimuli content and form, such as look&feel, pretext, or legitimacy of a message (RQ5.1). Additionally, we report whether the attacker of the simulated SE attack needs to *actively* interact with the target, for example, an instant message (IM) would usually require an *active* interaction from the attacker, while an email is commonly a one-off delivery with no further interaction.

Target parameters. We identify the (*personal*, *work-related*, *setting-related*) target parameters that have been included in the study as experiment's variables that relate to personal, work, or contextual characteristics of the targeted sample (RQ5.2). Here we only focus on target parameters because attack parameters are usually already implemented in the stimuli and their attributes, and are not used as treatments or control variables. Examples of *personal* parameters include demographics (age, gender, etc.) or personality traits (trust propensity, Big Five traits, etc.); *work-related* parameters - experience (years of service, training, etc.) or job position (junior, support, etc.); *setting-related* parameters - timing (e.g., number of days

between treatments) or device (e.g., mobile vs. desktop).

Perception. We indicate whether the study investigates any effects on perception, e.g., in terms of the presence of any *pre-attack* stimuli or *priming* operation prior to the delivery of deceptive stimuli (RQ5.3). An example of pre-attack stimuli can be sending an SNS request prior to attack stimuli delivery [45]; on the other hand, an example of priming is the influencing of participants with the notion of phishing before a phishing classification task [266].

Attention. We report when the study focuses on effects related to attention, i.e., as an (in)dependent variable (RQ5.4). To better characterize attention, we also extracted which types of (central) attention are engaged during the attacks since this influences Elaboration and conditions the final behavior. We identify two possible attention types as inferred from the experiment design: *exogenous*, when attention is a passive, transient, automatic, stimulus-driven process, and *endogenous*, when attention is a voluntary, sustained, goal-driven process in which the stimulus is aligned with the target's behavioral goals [57]. This distinction is relevant because, e.g., exogenous-driven attention can lead to a lower amount of cognitive resources used and the activation of heuristics [242]. For example, the use of *exogenous* attendance typically occurs in phishing susceptibility exercises at companies where the targeted employees are busy attending to their daily activities when the (unannounced) phishing email arrives. On the other hand, in a lab experiment where participants actively attend stimuli to, e.g., classify screenshots of phishing websites, attention is labeled as *endogenous*. When not enough information is provided, we label attention as *unknown*.

Elaboration. To provide an overview of the effects and interactions pertaining to Elaboration studied in the literature (RQ5.5), we identify experiment variables concerning the direct and indirect effects of stimuli and their attributes on Elaboration. Examples of such effects are the cognitive effort spent in processing a stimulus, measurements of reasons for a certain behavior, or the activation of cognitive biases [246, 371, 378]. We also include the activation of heuristics and anomalies, which is often linked to the manipulations of the artifact, e.g., the manipulation of the pretext of an email to reflect urgency or introducing misspellings in the text [42, 140]. It is worth noting that directly measuring the effects on Elaboration, such as the actual activation of heuristics and anomalies, could be particularly challenging as effects are difficult to isolate [91]. Therefore, their effects are often measured indirectly in relation to the outcome variables of SE susceptibility. To this end, we also investigate the studied indirect effects in the analysis along with the study types employed to measure them, as this allows us to get valuable insights into the state of empirical SE research.

Behavior. We identify the types of measured behavior (RQ5.6) to draw a picture of what are the different measurements of attack success investigated across the literature. Behaviors typically include clicks on links, submissions of information on bogus websites, or judgments of stimuli in classification tasks, for example, flagging legitimate/not legitimate websites or intention to reply/delete a message.

Features interactions. The collection of independent and dependent variables for each hypothesis/research question of a study allows us to examine the studied interactions between different variables mapped on the cognitive framework (RQ6). To this end, we record the

separate hypotheses and related variables along with the respective type (i.e., treatment, control, and outcome). An interaction is thus computed as an instance of two variables related to two distinct cognitive features on a per hypothesis basis, e.g., a hypothesis postulating that the usage of a persuasion technique in a message (*stimulus attribute*) has some effect on *elaboration*, and controls for subjects' age (*personal target parameters*), is counted as one interaction between *stimuli attributes* and *elaboration*, one between *stimuli attributes* and *personal target parameters* and one between *personal target parameters* and *elaboration*.

A detailed description of the analysis procedure is presented in Appendix B.1. The codebook used for the analysis and the raw dataset (including analyzed papers, empirical and cognitive features and exact values) are provided as supplementary material at `https://zenodo.org/record/8380243`.

4.4. Results

In this section, we present the analysis of our literature study. Results are presented by following the research questions specified in Section 4.3.1: we first present our findings with regard to empirical approaches adopted by SE studies (RQ1-4); next, we provide an overview of the cognitive features studied in the identified literature and a detailed analysis of each feature (RQ5.1-5.6). Finally, we analyze the interactions between cognitive features (RQ6).

An important consideration for results interpretation is that whenever a figure reports the number of papers ('`# papers`'), it should be read as "`the number of papers where [this feature] has been included/employed`" unless stated otherwise. This implies that the total sum of reported papers can be greater than the number of papers considered in the review, as a single paper can include more than one feature (e.g., modeling multiple covariates in a study). A comprehensive categorization of the literature sample is available as supplementary material at `https://zenodo.org/record/8380243`.

4.4.1. RQ1: What empirical methods have been adopted to study cognitive effects in the SE literature?

Fig. 4.3 shows the distribution of papers in our sample over the years. The first works appeared in 1996 and the number of publications has been constantly increasing across the years, with the exception of the last three, especially with regards to field experiments. Overall, the majority of studies are lab experiments (44%), followed by field experiments (42%), surveys (29%), and interviews (7%). Some papers report on more than one study type; for example, 19% of all field and lab experiments are complemented by a survey to capture additional variables and relevant factors needed for the testing of one or more hypotheses, such as participants' susceptibility awareness [209] or their reasoning patterns [216]. Interviews appear to be less common to study cognitive effects, albeit present throughout the whole period, whereas observational (e.g., retrospective) studies looking at cognitive effects are only rarely reported in the literature. In particular, we find only three observational studies that involve some kind of measurement of cognitive features. For example, one such work measured clicks from real phishing campaigns by studying data from taken-down phishing websites and correlating the behavior of users with the campaign attributes [138]. Another

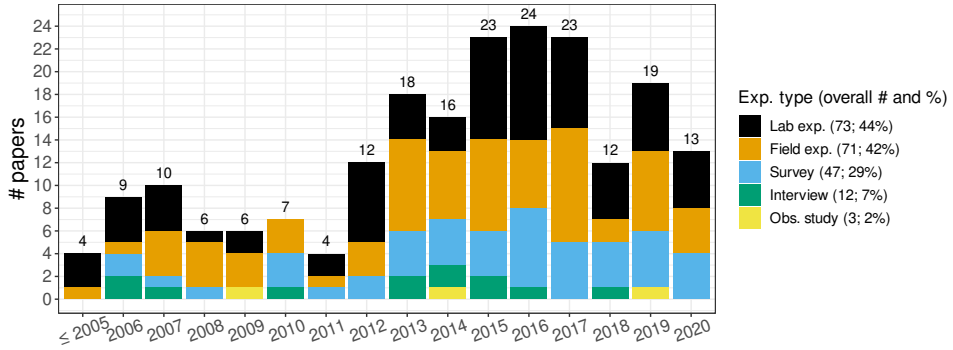


Figure 4.3: Distribution of papers by year across study types.

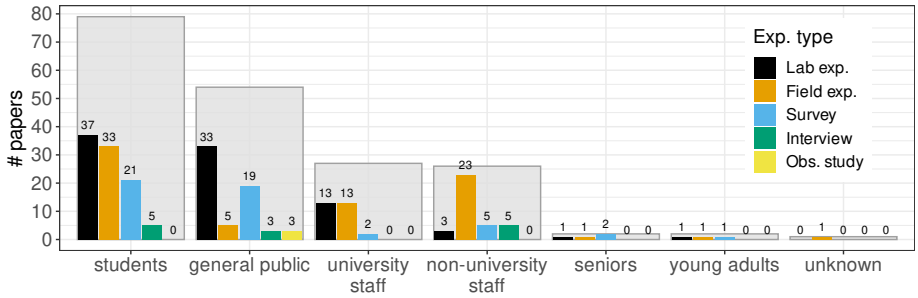


Figure 4.4: Distribution of papers by targeted population across study types.

study estimated actual and future clicks on links in real phishing campaigns as a function of persuasion techniques employed in the emails [347]. A distinguishing aspect of these works is that they measure *real* SE attacks and *real* user behavior, for which data is generally difficult to gather, explaining the relatively low number of such studies in the extant literature.

4.4.2. RQ2: Which subject populations have been considered to sample targets in empirical SE literature?

Fig. 4.4 provides an overview of the subject populations employed in empirical SE literature across study types. The figure shows that almost half of the studies involved student populations as participants (45% of papers), targeting mostly university students and, in two cases only, pre-college students [199, 325]. The second most frequent target population comprises users from the general public (33%), that is, subjects not sampled from any specific group (such as an institution). This category mainly consists of general Internet users and, in some cases, some very broad categories such as Facebook users [45, 357] or eBay users [164]. University staff (faculty and support) and non-university staff (company or institution employees) come in similar proportions (15% and 14%). Two studies involve a less common sample of participants: a study on phishing susceptibility of seniors [242] and a field experiment with older vs. young adults [209]. The unknown category contains two studies for

which it was not possible to determine the population type due to insufficient details in the experiment description (e.g., participants recruited with fliers around the campus, but no further detail is provided [383]). Section B.2 in Appendix B reports the distribution of sample sizes across the study types, and the supplementary material details the exact sample size for each reviewed study.

The prevalence of studies targeting the student population suggests that many experiments are carried out with convenience samples out of the pool of subjects available at universities. This indicates an overall under-representation of studies focusing on other target groups, such as employees in organizations or professionals active in different domains. University and non-university staff samples are almost equally represented; however, the former represents only one type of organization whereas the latter consists of a mix of companies and institutions operating across very different domains such as finance, construction/manufacturing, and NGOs. This is at odds with the observation that targeted attacks often aim at companies and institutions other than universities [39], suggesting that academic studies and experiments may be overall of only limited relevance for 'real world' attacks. For example, findings in [61] and [192] suggest that the effects of targeted attacks may vary substantially depending not only on subject characteristics but also on the domain in which the organization operates. Interestingly, experiments with general population samples are mostly lab experiments, while non-university staff are for the large part used as a subject pool in field experiments. This may depend on the effort needed to implement certain recruiting procedures (e.g., recruiting general public participants for a field experiment may be more difficult than for an online lab experiment), and the need to achieve a desired control of study variables (i.e., in field settings it may be unattainable to control specific factors such as workload or attention at the moment of the attack). On the other hand, some of these difficulties may be mitigated for experiments in organization settings, where the researcher may have access to fine-grained data to control, for example for stratified sampling, or measuring confounding variables.

Overall, we observe a general trend of recruiting subjects from the general public in lab experiments and non-university staff in field experiments. Conversely, students are largely recruited in both types of studies, with potential limitations on the external validity of the associated findings. Therefore, the problem of characterizing the effects of SE attacks on company and organization employees, and across domains, remains open. Further, highly vulnerable categories such as senior citizens and youngsters remain widely understudied.

4.4.3. RQ3: What types of artifacts have been considered for the delivery of social engineering attacks in empirical SE literature?

Fig. 4.5 shows the prevalence of different types of stimuli across study types. Emails are the most commonly employed stimuli followed by websites and, to a lesser extent, SNS. URLs only and voice calls are also studied, but much less prevalent. Other stimuli encompass Wi-Fi access points and instant messages [183], pop-ups [242], physical media (brochures, mail, affixed QR codes) [174, 354, 376] and in person deception [55]. The high popularity of attacks with emails and websites is not surprising, given their popularity as attack vectors. On the other hand, attacks conveyed by media other than email and websites are widely underrepresented despite these being increasingly often reported in the wild [39]; examples are

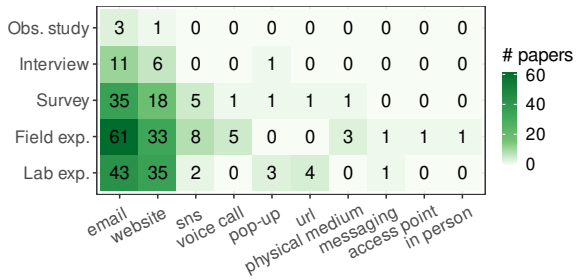


Figure 4.5: Distribution of papers with respect to stimuli and study types (SNS: Social Networking Site).

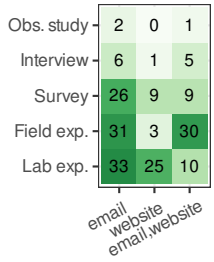


Figure 4.6: Zoom-in on combinations of email and website stimuli types.

4

deception over voice [43] or Stuxnet-like attacks over USB drives [197], and more in general multi-step-multi-media attacks, such as reverse SE on LinkedIn [15] or lateral movement attacks in organizations [77]. Studies reproducing these attack scenarios are largely not yet reported in the literature.

As emails and websites are tightly linked and generally part of the same attack procedure (e.g., email with a URL linking to a counterfeit login interface), Fig. 4.6 offers a breakdown of these dimensions. Studies reported under each label are studies that employ that stimulus type (e.g., URL) but not the other (e.g., website). Studies employing both are reported as *email,website*. We can observe that the majority of experiments investigate these stimuli individually. This implies that most studies assume that phishing attacks are successful when the target executes one action only (e.g., click a link or open an attachment) [88, 380]. On the other hand, this is generally not the case in reality [120]. Therefore, these studies may not accurately capture real victimization rates. For example, especially in field studies, users may want to preview a URL they detect as phishing out of curiosity, without the intention to input their credentials on the landing webpage [119]. Further, users are known to commonly engage in multi-modal communications (e.g., voice and text, email, SNS, instant messaging) for both personal and professional communications. These happen across multiple devices, such as personal computers and portable devices all with their own user interfaces (that condition how stimuli are consumed, and how deception takes place). These dynamics stress the need to account for multi-step-multi-modal scenarios for future experiments in this area.

Nonetheless, we find a number of studies featuring email and website combinations (and possibly other stimuli on top of those[4]). We observe that the majority of these are field experiments, for example simulating phishing campaigns with a website asking for some information. Interestingly, there are only a few lab experiments in our sample that reproduce two-step SE attacks, e.g. [17, 199, 381]; the sheer majority of lab experiments feed participants with one stimulus at a time, often in a static fashion, such as screenshots of emails, where no interaction is possible. This signals a tendency to prefer one-step attacks in

[4]Among the few works that did combine email, website, and other stimuli, Workman et al. [375–377] utilized a combination of email, website, voice call (and mail) but without providing details on the specific behavior(s) being measured.

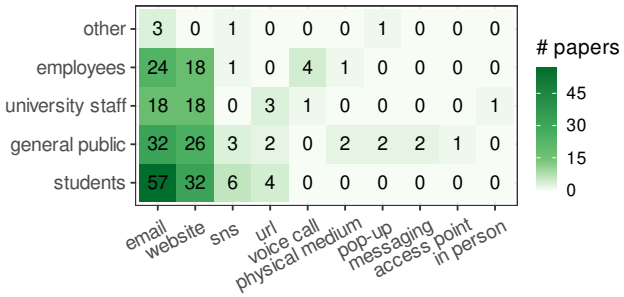


Figure 4.7: Distribution of papers with respect to stimuli types and targeted populations.

laboratory settings with either emails or websites compared to other more realistic and complete experimental setups. This highlights an open opportunity to explore cognitive effects in multi-step attacks with more simulations in laboratory settings; some researchers have already moved in this direction [105, 257, 322]. From Fig. 4.6, we can observe only three field experiments that use websites in combination with other stimuli types; these studies use access points [183], SNS [317], and QR codes [354].

Fig. 4.7 shows the distribution of papers across stimuli types and targeted populations. Emails and websites have been studied with all types of target populations, while other stimuli types are not evenly represented across populations: for example, voice calls were exclusively utilized with university and non-university staff, and social networks mostly with students and the general public. There is only one study where SNS were employed to simulate an attack against company employees by impersonating a fake employee to infiltrate closed groups on LinkedIn and later post a malicious link [317]. Various media display different information to the user, their perceptions, expectation and interaction change accordingly. Therefore, these gaps fall short with the increasing adoption of SNS, messaging apps, and even QR codes across all sectors of society (e.g., QR codes at bus stops, or on pub tables to access menus), emphasizing the importance of expanding SE empirical research to cover as many diversified population samples with as much attack surface as possible, for example, in terms of stimuli types and sequential attack stages. To this end, frameworks able to capture the (cognitive) processes triggered by SE attacks, as the one presented in [57], can help define a coherent and consistent account of possible interactions between multi-stage stimuli and help identify the relevant target population and methodological choices to employ for their investigation.

4.4.4. RQ4: To what extent are SE artifacts tailored to the experiment subjects in empirical SE literature?

Fig. 4.8 shows the distribution of attack targetization over study types. The studied attacks are mostly targeted against generic populations (68% of papers) and against specific populations (32%). Only four papers study attacks against specific individuals (2.4%), for example, by investigating the effects of increasing degrees of targetization [337] or of training against spear-phishing attacks [62]. By contrast, recent attacks already show signs of automated tailoring, such as automatic detection of the affiliated company based on the domain in the

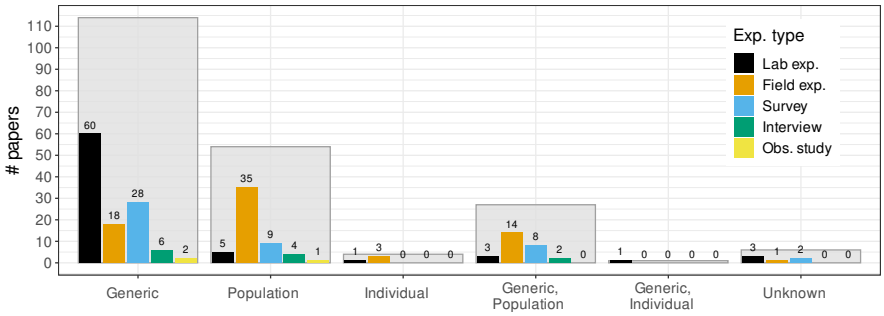


Figure 4.8: Distribution of papers by attack targetization across study types. Generic, Population, Individual, Unknown.

user's email address and integration of that company's logo into a fraudulent webpage [326]. On this same line, a recent research work presents a toolkit to leverage OSINT information on targets to deliver large-scale spear phishing training campaigns [281]. This positions the current state of empirical SE well behind the future of scalable targeted attacks against which current defensive strategies may need to be adapted. Surveys and interviews capable of providing qualitative insights on targeted effects are especially lacking.

From Fig. 4.8, it also emerges that a number of studies consider *both* 'general' and 'population'-targeted attacks (16% of papers), showing that attack targetization is responsible for large changes in expected success rates. For example, Holm at al. [150] report a fourfold increase in attack success rate when the pretext is tailored to the subject population. A number of other studies provide similar insights [18, 19, 375, 377], suggesting that targetization is an important variable to account for in SE studies. This contrasts with the fact that many studies do not gauge the level of targetization of the used stimuli. Hence, it often becomes impractical to compare the results between apparently similar experiments with similar populations, but with differently adapted stimuli [132]. To address this, future studies could benefit from a consistent accounting of the adaptation degree between stimuli, the targeted population, and, where possible, the target context. With respect to the framework in Fig. 3.1, this translates to determining the degree of matching between target and attack parameters and accounts for such a degree during artifact construction.

4.4.5. RQ5: Which cognitive features of SE attacks have been tested empirically in the SE literature, and in which experimental settings?

Fig. 4.9 presents an overview of the cognitive features studied in the identified literature, and their employment in the respective experimental designs as treatment, control, or outcome variables; features only indirectly included in a study design are reported as 'indirect'. Unsurprisingly, the most studied feature is behavior, mainly as an outcome variable (in 96% of papers) and/or as a control variable (25%; again, note that a single variable may be used in different ways within the same paper). Personal target parameters are the second most studied feature (14% as treatment, 59% as control, and 21% as outcome) followed by stimulus attributes (29% as treatment, 58% as control, and one instance as outcome); by contrast,

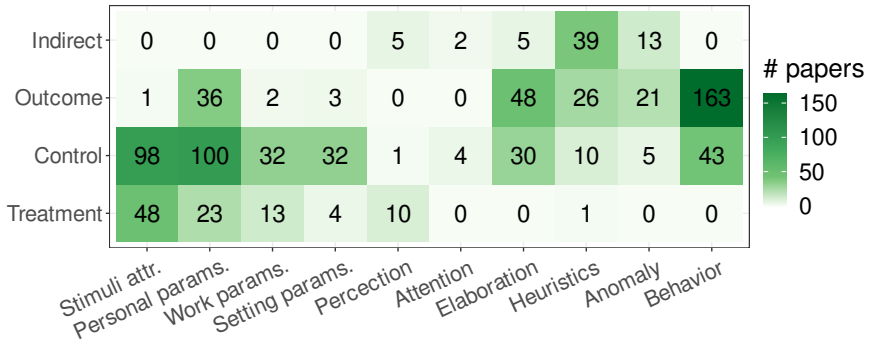


Figure 4.9: Distribution of framework features in the experiments and their employment as treatment, control, and outcome variables; variables only included indirectly in a study design are reported as 'indirect' in the figure.

only a few papers consider perception (6% as treatment, 3% indirect and one as control) and attention (2% control and 1% indirect). This suggests that extant research tends to focus more on the effect of subject characteristics on phishing than on contextual factors affecting the subject at (or around) the time of the attack.

All target parameters (*personal target parameters*, *work-related target parameters*, *setting-related target parameters*) are for the largest part used as controls. However, *personal target parameters* and *work-related target parameters* have also been instrumented as treatments, often in the form of anti-phishing training/awareness in either personal or work-related contexts. *Personal target parameters* have also been instrumented as outcome variables in studies interested in situational variables (such as perceived susceptibility, awareness, risk) and in how these variables are influenced by other factors, e.g. [20, 244]. *Setting-related target parameters* have been almost exclusively used as control variables; when used as treatments, they were employed in the form of incentives provided to the participants [206, 245, 316, 388]. Overall, we find that the extant literature tends to focus on the personal characteristics of the subjects, overseeing the setting in which the attack takes place.

Perception has been mainly 'manipulated' with treatments aimed at 'priming' participants (e.g., [45, 65]) or indirectly triggering the activation of generic vs. specific percepts with highly contextualized stimuli (e.g., [125, 150]). We find that attention is the least investigated feature in empirical SE: two studies indirectly manipulated attention [361, 364] and four studies used attention as a control in relation to the outcome variable, e.g. [242, 372].

Elaboration, Heuristic, and Anomaly are generally studied in terms of outcome and indirect effects and are less frequently employed as control variables. Attempts to measure Elaboration features include mostly cognitive effort and elicitation of reasons for behavior (e.g., [244, 354]). Heuristics and Anomalies features are predominantly studied as indirect effects of cognitive biases and anomalies [42], and as outcome variables in terms of heuristics and trust indicators [88, 361].

The overview of the status of empirical SE research reported above indicates that there might

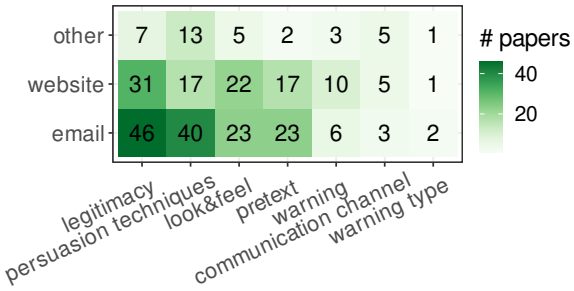


Figure 4.10: Distribution of papers with respect to stimuli attributes and stimuli types.

be certain 'boundaries' with respect to what is being measured and what is possible to measure with the available techniques. For example, investigating the effects of stimuli attributes and target parameters on behavior is highly relevant, particularly for more advanced and targeted attacks, but somewhat limited by the uncertainty of indirect measurement methods; on the other hand, directly measuring and manipulating Elaboration-related features would be invaluable but so far infeasible except in very narrow applications: some cognitive processes simply cannot be measured till technologies, such as brain implants, are made available [282].

RQ5.1: What stimuli attributes have been investigated?

Fig. 4.10 shows the distribution of stimuli attributes across stimuli types. Whereas a large body of literature evaluates the effect of legitimacy, persuasion techniques, look&feel characteristics (i.e., layout, design, logos, writing style, etc.) and pretexts, the effect of warnings and the means by which the stimulus is delivered to the target (communication channel) appear to be far less developed. Active interactions across multiple stimuli between attacker and target were utilized, in our sample, in only nine studies, for example, in voice call pretexting [4, 51, 55, 375, 377] or social network interactions [355, 357, 359] (category 'other' in Fig. 4.10). The effects of these *active* interactions on cognition and attack success were, however, not thoroughly investigated, leaving ample room for further studies, given also their saliency in recent attacks [15].

The mapping of attributes on the stimuli types shows that persuasion techniques (i.e., the exploitation of certain human cognitive biases), have been covered in the literature across all stimuli types. The legitimacy attribute, i.e. the stimulus being legitimate or non-legitimate, is often instrumented in an experiment as a control variable by collecting samples of real deception attempts and legitimate communications, and by administering them to the participants to assess their ability to judge the stimulus legitimacy [184]. This method is a popular and easy way to estimate the susceptibility to SE attacks of a given population (cf. Section 4.4.5). However, the legitimacy attribute is almost entirely investigated for emails and websites, but very seldom for other types of stimuli. Similarly, the look&feel of other types of stimuli is also under-explored, for example, the case of instant messaging apps or SNSs is certainly worth investigating deeper given their widespread and the great potential for misuse [324]. There is thus a lack of studies exploring the effects on elaboration and behavior of otherwise

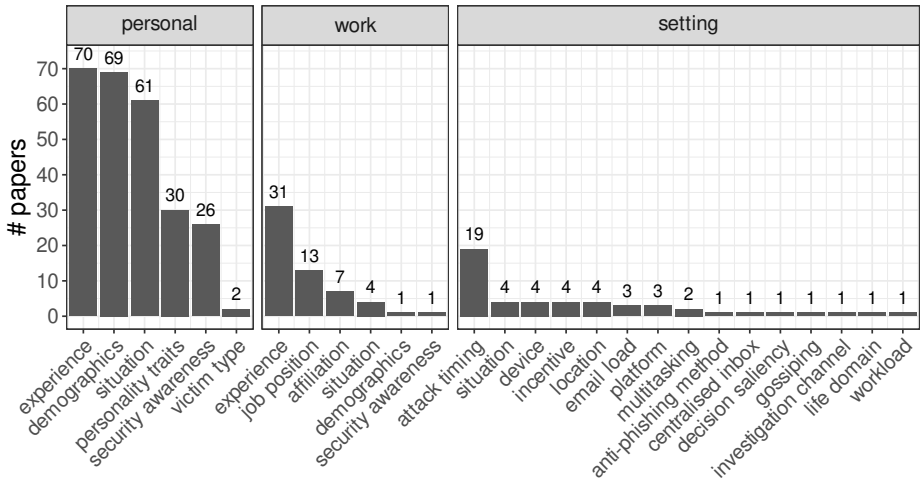


Figure 4.11: Distribution of papers by parameter type and category.

commonly investigated stimuli attributes (excluding persuasion techniques) on SNS, instant messages, or voice calls. As mentioned in Section 4.4.3, other stimuli types represent a rich avenue for new and sophisticated attacks that can capitalize on the diversity of media and mix personal and professional life domains.

A number of studies in our sample investigated defensive mechanisms against SE attacks as part of their experiments (e.g., [1, 95, 104, 384]). Such studies often involved testing warning messages (as shown in Fig. 4.10) aimed at preventing user deception. However, the *type* of warning messages, i.e. the different designs and contents of such messages, was seldom considered. This is at odds with the unclear effectiveness of many standard warning messages reported in the literature [96, 362, 381]. Investigating warning types is important because effective warnings can help users make the right decision when subject to SE attacks. The experimentation with new and non-intrusive warnings and interventions (e.g., *nudges*) has been recently highlighted as a valuable opportunity to improve current defense techniques [53, 58, 121].

RQ5.2: What target and contextual characteristics have been investigated?

Fig. 4.11 shows the distribution of the studied target parameters, namely *personal target parameters*, *work-related target parameters* and *setting-related target parameters*. Parameters have been grouped in logical categories (we refer the interested reader to the codebook provided in the supplementary material at `https://zenodo.org/record/8380243`). It is worth noting that the same name has been used to denote certain categories related to different types of target parameters although they represent different parameters. For instance, demographics in personal parameters includes age, education, etc. while in work-related parameters it includes salary (for more details see also Appendix B.1).

It is immediate to observe in Figure 4.11 that the sheer majority of investigated parameters

fall into *personal target parameters*. Experience and demographics are the most commonly studied personal parameters (40% and 38% of studies respectively), as they are traditionally included in experiments with humans, e.g., age, gender, or level of security knowledge and training. The third most frequent parameter category, situation, encompasses short-term cognition factors that are often situation dependent, such as perceptions of risk [130] or self-efficacy [386] in a task. While such factors represent constructs related to the personal subjective dimension of a given situation, these are not to be confused with *setting-related target parameters* which regard the *contextual* dimension of the circumstances of the experiment, such as the environment [318], timing [97] or the concurrent activities of the participants [361]. Other personality traits represent long-term factors, such as propensity to trust or the *Big Five Inventory* (BFI) personality measures [172]. A number of personal subject parameters show consistent negative effects on SE susceptibility, such as experience [380] and knowledge [145, 364], whereas others show mixed or no effects (e.g., age [266, 375], gender [266]). Subject parameters such as curiosity and commitment are often reported to have a positive effect (i.e., they *increase* attack susceptibility) [239, 375].

Work-related target parameters mostly consider subjects' professional experience (15%) such as years of service and security training in the working environment [177, 369]. Job position (e.g., student, professor, management, support staff) is oftentimes used as a proxy for familiarity with the overall organization context [61]; however, this may introduce errors for newly hired professionals in senior positions [54]. Only one study in our sample evaluated the effects of subject parameters across different organizations (yielding mixed outcomes) [61].

The most common setting target parameter is attack timing (10%), as employed in anti-phishing training studies. Only three experiments (2%) measured the effect of different devices on which the stimuli is received: two field [358, 359] and one lab experiment [237]; of these, the two field experiments reveal a significant positive effect (i.e., increasing the attack success rate) of using a smartphone as opposed to a desktop environment. The lab experiment found no significant differences, albeit we note that it was carried out in a controlled office environment, which is far off from the ordinary context that subjects experience in field experiments [237], making the two results hardly comparable. Nonetheless, the effects of contextual factors (while relatively unstudied) may be decisive on the outcome of SE attacks, as highlighted in [132].

Overall, our first and most important observation is the net tendency of the state-of-the-art to favor *personal target parameters* and *work-related target parameters* in their investigations. Whereas individual personal and work-related factors are undoubtedly important in affecting the outcome of SE attacks, setting parameters are largely dismissed as of secondary importance both in terms of the number of studies that investigate such factors and their variability across the studies; indeed, hardly any setting-related factor is constantly considered in the literature, despite their reported importance [132]. For example, only four studies considered smartphone usage (grouped under `device` in Fig. 4.11), which is surprising considering the popularity of mobile phones and the usability constraints they introduce both at the interface level [358] and at the contextual level [316, 358] (e.g., multitasking while on the go).

The lack of studies considering setting-related factors may be partly due to the inherent dif-

ficulty to control for such variables. Nevertheless, some studies were able to measure some aspects of the target's context, such as workload [166] or email load [371], or the life domains the targets are sensible to [209]. Studies from other disciplines, such as social sciences, can provide valuable methodological insights. For example, a study on clinical reasoning asked participants to watch video-recorded clinical encounters (treated with patient contextual factors related to emotional volatility and language proficiency) and produce a diagnosis; the investigators then measured the effects of such factors on diagnosis accuracy [226]. This methodology can be conveniently adapted to, e.g., laboratory experiments in SE where participants are given a framing scenario for a task, but with modified contexts. This example illustrates that future experiments may benefit from new techniques or techniques adapted from other disciplines to control promising contextual factors, such as the operational setting in an organization or the shared vs. individual domain of current activities [132].

RQ5.3: What effects on perception have been investigated?

Perception translates stimuli into percepts, which represent the mental result or product of perceiving. Depending on how well a stimulus aligns with the target's context, perception can load more specific percepts in a subject's working memory [57]. In this respect, an attacker's attempt to condition perception with *pre-attack* and *priming* operations before the delivery of deceptive stimuli might play a critical role in the success of SE attacks (cf. Chapter 3). However, despite the relevance of priming in SE, only a few papers in our collection have studied its effect either for defense [65, 129, 131, 160, 170, 174, 266, 268, 270, 271] or attack [45], and overall the effect of priming still remains unclear. For example, Benenson et al. [45] do not find significant effects of priming on attack success (i.e., sending SNS friend requests before the actual attack). Many studies on priming in defensive scenarios [65, 129, 266, 268, 270, 271] do find that priming has a significant positive effect, while other studies [65, 131, 160, 174] report no significant effect of priming subjects before similar phishing classification tasks. Priming can also have an impact on the amount of cognitive effort subjects employ in their defensive decisions [266]. Whether the opposite is true in an attack scenario is still an open question. Albeit the considerable uncertainty around the effects on perception, related attacks can represent an untapped extension of the attack surface exploitable by the attackers. Overall, priming for attack scenarios calls for further experimentation; relevant techniques may be borrowed from the field of social psychology and cognitive sciences such as social stereotypes [90] and subliminal triggers of affective reactions [374].

Only a few studies investigated the specificity of target-related information and contextualization in phishing attacks [125, 150, 154, 184]. We find that most studies employ only a 'general' perception, whereas only two studies design specific attacks likely to trigger highly-specific percepts in their targets [62, 337]. Overall, a detailed account of perceptual mechanisms and their effects in the context of SE is still inconclusive. Once again, methods applied on perceptual and memory-based influences [302] may be used, for example, to evaluate the performance in phishing classification tasks of inexperienced users, or subjects acting under time constraints.

Furthermore, whereas the suggested social and cognitive sciences literature focuses on information-rich media, such as in-person or verbal communication, the SE literature has focused

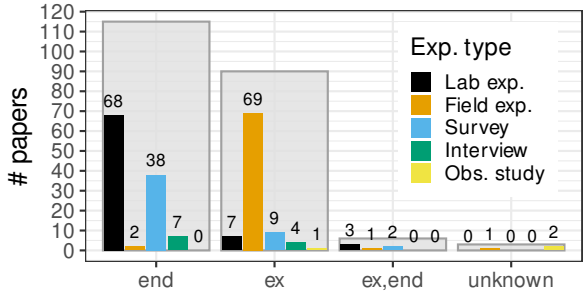


Figure 4.12: Distribution of papers by attention type across study types.

mainly on written communication forms, i.e., emails and websites. This suggests that effects on perception may be particularly relevant in verbal communication, i.e., vishing attacks as well as socially rich written communicative such as SNS or lateral movement attacks in organizations. For example, caller ID spoofing and internal/familiar entity impersonation can significantly increase an attacker's success over voice calls [339], as also exemplified by recent vishing attacks in political cases [43] and recent mass social security scams involving expats [102]. Similarly, the involved perceptual mechanisms in written attacks still have the potential to make scams more convincing, such as friends' recommendations on Facebook [158] or the specificity of tailored scenarios [15]. We argue that this represents an opportunity to define a new research line to test and address new, unconventional forms of attacks involving perception.

RQ5.4: What effects on attention have been investigated?

Attention modulates the conscious elaboration of stimuli, where the two types of central attention considered here (*exogenous* and *endogenous*) influence the tendency of Elaboration to occur heuristically or consciously (cf. Chapter 3). From Fig. 4.9, we can observe that only six works studied the effects on attention: five studies [242, 361, 364, 372, 379] investigated the effect of attention type (endogenous, exogenous) and one measured the level of (endogenous) attention [287]. Wang et al. [364] and Wright et al. [379] employed surveys to find out, retrospectively, which attention type has been engaged and showed a significant correlation between attention type and phishing susceptibility (with exogenous attention leading to higher deception rates). On the other hand, only Morgan et al. [242] manipulated the attention type by setting the experiment to (indirectly) set participants' attention to be endogenous or exogenous to the specific task. This study supports the effect of attention on lowering the amount of cognitive resources deployment, with exogenous attention leading to higher chances of heuristic processing. Results in [287] indicate a strong positive correlation between the ability to exercise sustained attention (closely related to endogenous attention [306]) and the ability to correctly classify phishing websites. These studies signal a trend to associate exogenous-like attention with higher attack success rates; conversely, studies where (the degree of) endogenous attention is explicitly evaluated are still lacking.

Fig. 4.12 presents a breakdown of attention types with respect to study type. Laboratory experiments (and surveys) almost always imply the use of endogenous attention, while field

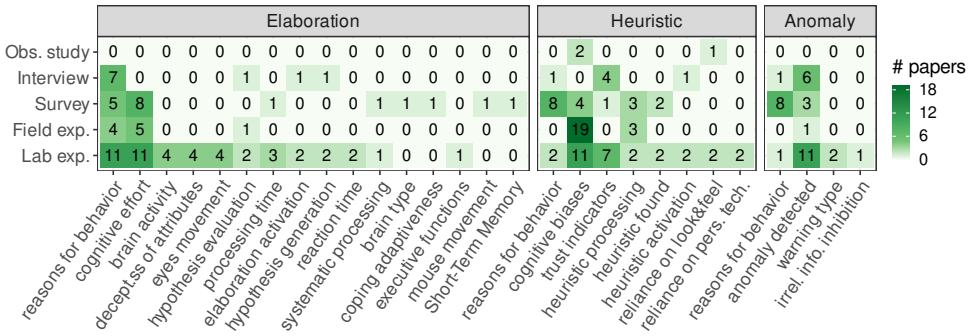


Figure 4.13: Distribution of papers by features related to Elaboration, Heuristic, and Anomaly w.r.t. study types.

4

experiments employ exogenous attention. Studies marked as '*ex,end*' involve both types (e.g., [4, 151, 166]) or control for attention type [242, 372].

Overall, these observations indicate that attention can play a decisive role in the outcome of a SE attack. This is particularly topical as practices such as the introduction of *Bring Your Own Device* (BYOD) to work settings, or the employment of mobile devices for work-related and personal tasks alike may significantly affect the outcome of an attempted attack. As these effects are widely understudied, this opens a large research gap calling for new studies aimed at understanding how attention is affected by contextual factors of the target, such as device type or physical environment, and the effects of this on attack susceptibility and possible countermeasures. Nevertheless, only a handful of studies reproduced scenarios where attention can be reasonably manipulated (as in [242]) and related to, e.g., matching of target parameters (as in [358]). Techniques to manage attention can be adopted from other fields, such as parallel recognition tasks from cognitive sciences [207]. Being able to determine and control the kind of attention deployed during laboratory or field experiments would allow an unprecedented step forward in the comprehension of the mechanisms responsible for conscious and unconscious determinants of 'right' and 'wrong' decisions during SE attacks. Furthermore, methods for defense would benefit from these advancements by, for example, developing interfaces able to nudge attention to spot anomalies without negatively affecting the usability of applications [121].

RQ5.5: What effects on elaboration have been investigated?

To characterize and determine the boundaries of the effects on Elaboration that have been investigated and to shed light on what has not been investigated in relation to the experimental constraints, we provide an overview of the effects pertaining to *Elaboration*, *Heuristics* and *Anomalies* with respect to study type in Fig. 4.13. Across study types, the most investigated features are cognitive biases, reasons for the adopted behavior, and cognitive effort. From the figure, we observe that lab experiments are the preferred method to investigate features concerning elaboration. For example, a phishing classification task by Parsons et al. [267] included open questions about participants' reasoning for decision making to develop a framework on user intention and actual behavior; Nicholson et al. [249] tested anomaly detection

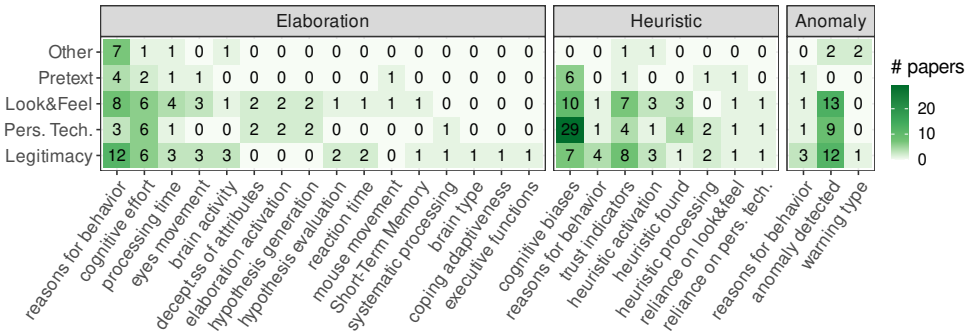


Figure 4.14: Distribution of papers by features related to Elaboration, Heuristic, and Anomaly w.r.t. stimuli attributes.

4

with saliency nudges as treatments in an online lab task. Similarly, but in a more elaborate lab setup, Hale et al. [140] explored heuristic activation and anomaly detection. Field experiments have also been adopted to study elaboration features, especially concerning *Heuristics*: Williams et al. [371] investigated the triggering of cognitive biases by means of persuasion techniques and reconstructed reasons to respond to or report a phishing email. In another phishing simulation, Caputo et al. [66] measured reasons for behavior and cognitive effort by interviewing "*clickers and non-clickers*" after the fact. Interestingly, both studies reveal that subjects mentioned (correct and incorrect) strategies to quickly make a decision. Standalone surveys and interviews are nonetheless employed to investigate some features of elaboration such as heuristics usage, e.g. [360, 361], trust indicators or anomaly detection, e.g. [165, 275]. Measurements of effects on elaboration and their relations to the outcomes of an experiment can be generally considered as indirect, given that such measurements result from conscious, after-the-fact elicitation that may not always be accurate: people might have limits in their *motivation* to report mental content of which they are aware; limits in their *opportunity*, given the *circumstances* of a measurement; as well as limits in their *ability* and *awareness* (inaccessible mental content) [254]. Nonetheless, more direct measurements have also been adopted in certain cases. For instance, cognitive effort has been measured as a function of time [266] or by means of eye-tracking devices [246], which are employed to identify visual focus areas during the elaboration [234]. Additionally, several attempts to directly measure brain activity have been carried out with, e.g., fMRI during phishing classification tasks [247, 248, 346], where, for example, brain areas responsible for executive functions are more active when subjects are explicitly asked to evaluate phishing stimuli, than when asked to just look at stimuli without judging [248]. However, performing such measurements is challenging as effects remain difficult to isolate [364].

Fig. 4.14 shows the interactions between features of elaboration and stimuli attributes. Persuasion techniques are often implemented in the stimuli to trigger cognitive biases in the targets. The most investigated cognitive biases are Scarcity/Urgency, Authority, and Liking, whose triggering is usually inferred from the outcomes of a simulated attack. With this approach, however, it is difficult to control confounding effects stemming from each individual's characteristics and context (i.e., target parameters). To mitigate the resulting

Figure 4.15: Distribution of papers by behavior w.r.t. study types.

4

uncertainty, several studies implemented additional measurements; for example, Parsons et al. [265] tested persuasion techniques and controlled for impulsivity as a proxy of a subject's propensity for heuristic or systematic decision-making. Similarly, Vishwanath et al. [359] explicitly asked subjects for the "*heuristics they generally use*" in the designed scenario and which stimulus cues (i.e., picture and number of friends on a Facebook page) they pay attention to. Look&feel is often related to trust indicators in heuristics; for instance, the lab experiment reported in [208] investigates how a webpage content and URL can influence the process of consciously evaluating whether something is deceptive and which trust indicators subjects rely on when classifying phishing. Look&feel along with pretext or other stimuli attributes have also been studied in combination with persuasion techniques. For example, one study [42] relates look&feel characteristics and the subjects' detection of 'anomalies' in the same hypothesis. Similarly, the pretext used in the attack may be sometimes related to persuasion techniques and consequently to cognitive biases [125]. In accordance with previous literature [320], we observe that the effects of the pretext on elaboration are less explored (especially heuristic and anomaly activation) in spite of the pretext being often regarded as an important explanatory variable in real and simulated attacks [132, 209].

Nonetheless, nearly all studies concerning effects on elaboration rely on conscious elicitation to measure such effects. To overcome the general limitations of conscious elicitation of elaboration, SE research may need to resort to implicit measurement procedures (i.e., that do not rely entirely on conscious elicitation) [254], or develop methods able to measure features when the processing phase is still 'hot', that is, immediately after a certain action is performed [121] (e.g., instant feedback [322]). The present state of knowledge in social cognition provides a plethora of implicit measurement methods and research evidence as summarized in [254], with applications spanning from organizational [342] to consumer research [294]. The applicability and adaptation of such techniques to the SE domain remain, however, an open question.

RQ5.6: What types of behavior have been investigated?

Fig. 4.15 presents the distribution of behaviors considered in the extant literature across study types. Classification of stimuli is the most studied behavior (e.g., phishing vs. legit

emails) followed by clicking a link and submission of sensitive information such as credentials. Whereas these are usually considered by themselves as proxies for deception success (e.g., [66, 182, 364]), in general, a single action may not necessarily lead to a security impact; rather, the impact realization may depend on the resources of the attacker and the type of system (e.g., determining the success or failure of a *drive-by-download* attack [189]), or on the characteristics of subsequent stimuli (e.g., a badly cloned website). Moreover, visiting a malicious website does not currently represent a high-security risk due to the countermeasures employed by most OS and browsers (such as the wide spread of Address Space Layout Randomization, auto-updates, and phasing out vulnerable technologies, e.g., Adobe Flash or Java), eventually leaving macros-enabled documents as the preferred attack 'click-vector' [80, 112, 351]. Yet, the assumption that a click corresponds to a security impact is oftentimes (explicitly) made: "*clicking […] deploys malware and opens virtual backdoors*" [356], "*the click of an email link can take users to a fake site requesting login information*" [154] or "*expose the organization to a network of hackers*" [166]. None of these works, however, modeled or measured the actual compromise, information submission, or exploitation by means of, e.g., submission forms or executables [120, 192][5]. Whereas relevant, the implicitly assumed threat model diverges significantly from that of a 'regular' attacker (see, e.g., [47]), with unclear implications on the realism of the simulated attack procedure (including the implementation of the pretext). An alternative to malware infection via link clicks is by means of email attachments; yet, surprisingly, we find very few studies of this type. For example, only two studies employ archive files as attachments [169, 322], other two PDF files with links [358, 360] and one with an HTML attachment [96], whereas we find no study employing MS Office documents, despite their importance as delivery vector of malware [269].

A number of studies do distinguish link clicks from submission of information (e.g., [150, 192]), submissions only (e.g., [61, 145]) or consider the opening of artifacts and executables (e.g., attachments [360] or downloaded files [150]) as measures of 'success'. Generally, experiments considering two-stage scenarios (e.g., phishing email and a subsequent landing webpage, see Section 4.4.3) report lower success rates than one-stage studies [320]. This suggests that real phishing success rates may be lower than otherwise reported by studies simulating only one attack stage. Finally, a few studies recorded and investigated the act of reporting SE artifacts to, e.g., IT departments [66, 96, 133, 378], despite reports being a valuable prevention and mitigation method [56].

The breakdown by study types shows that only a few studies investigate behavior using surveys or interviews. For instance, interviews have been used to correlate classification tasks, clicks, and responses with other features such as reasons for a given behavior, indicators of trust in stimuli, and identified anomalies (e.g., [66, 99, 380]). Interestingly, two studies report remarkable differences between *intention* to click and *actual* clicks, finding rates in the latter higher than in the former [151, 209]. However, surveys and interviews often do not consider other relevant behaviors, such as credential submissions and opening/executing artifacts. Neither do lab experiments, except a small user study with attachment-like artifacts [322]. Nevertheless, such experiments can be valuable instruments to capture user perception or decision-making, e.g., trust certain file types, enable macros in document files,

[5] This observation can be extended to past works as well where, e.g., browser exploitation was still relevant but its realistic impact on security (e.g., [189]) might be unclear.

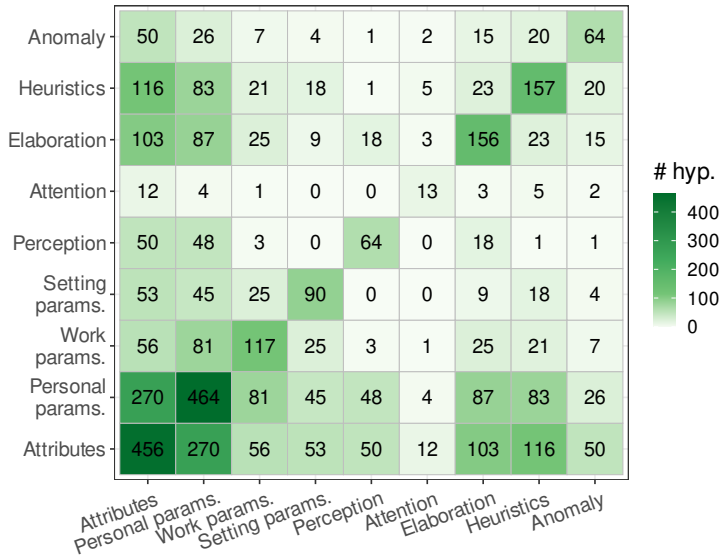


Figure 4.16: Studied interactions between SE cognitive features.

or report attacks to IT departments. As an example, a survey employing phishing emails with and without attachments reported that the *presence* of a file influences suspicion and heuristic processing in the subject, and ultimately conditions the attack outcomes [360]. It is thus important to foster investigations able to reproduce and measure such scenarios (e.g., with attachments) to get insights on what drives such risky behaviors and to devise suitable un-intrusive methods to guide the user in the right decision for such cases.

4.4.6. RQ6: What interactions between cognitive features have been studied in empirical SE literature?

Fig. 4.16 reports the distribution of cognitive features that have been studied together, as defined in the hypotheses of the sampled literature. It is worth noting that a data point in this figure is a hypothesis evaluated in a paper, as opposed to a paper; accordingly the diagonal reports the number of hypotheses that consider the respective variable. Interactions between stimuli attributes and target personal parameters, as well as personal parameters and work parameters, seem to be commonly explored in the literature. For example, pretext and persuasion techniques have been investigated across several hypotheses together with gender and personality traits [12, 125], or with job position and type of anti-phishing training [170, 240]. Similarly, attributes and *personal target parameters*' effects on *Elaboration* and *Heuristics* are also commonly explored (reflecting stimuli attributes and personal characteristics being the typical means researcher employ to measure the effects on *Elaboration* [356, 371], as also discussed in Section 4.4.5).

Studies investigating the interactions between perception and other cognitive features are, in general, rare (cf. Section 4.4.5). The only interactions that have been studied are with the

stimulus attributes (e.g., [174]), *personal target parameters* (e.g., [129, 266]) and to a lesser degree with *Elaboration* (e.g., [268]) and *work-related target parameters* (e.g., [170]). Perception can be influential in SE attacks that capitalize on exploiting the trust and expectations subjects place in certain media or platforms that are widely accepted to be trustworthy. This is the case, for instance, for the attack delivered through the LinkedIn platform described in [15].

Setting-related parameters are also, perhaps more surprisingly, seldom considered in relation to other variables. Notably, no hypothesis in our sample considered the effect of target setting parameters on attention and perception values, and very few evaluated setting-related parameters' effects on *Elaboration* (e.g., [159, 245, 316]). Similarly, we find large gaps between cognitive processing at *Elaboration* and attention levels (e.g., [63, 287]), and little work explicitly studying the relation between anomalies and heuristics with *Elaboration* dynamics (e.g, [145, 360]). Yet, insights on interaction effects with *Elaboration* and other features can hold valuable implications for the development of more effective anti-phishing education efforts. For example, increased elaboration and attention to incoming email messages may not be an effective strategy in making the right decisions and, perhaps, it is better to teach users to rely on only a few key elements in the message (e.g., the actual address) [145]. On the same wave, studying the interactions with attention can increase the understanding of SE attack processes, especially more complex processes which are otherwise more difficult to measure or reproduce. For example, one can attribute an influential role to attention in highly interactive and fast-paced attacks, such as vishing attacks. A notable case is an OSINT investigation in a high-profile political assassination attempt [43] where the attacker attempts to deceive his targets into revealing information by means of authoritative impersonation (spoofing caller ID) and overloading the targets' attention with several contextual details. These examples underline the relevance of gaps in the map of Fig. 4.16, which clearly shows that the extant literature has focused on a rather narrow research space.

4.5. Discussion

In the previous section, we analyzed the collected studies on the various dimensions represented by our research questions, as outlined in Fig. 4.1. In this section, we summarize our findings and identify gaps in the literature and promising directions for future work.

Gap between real attacks and attacks simulated in the studies

The main conclusion from our analysis is that real(-istic) attacks are only partially reflected in the experimental setups employed by a large portion of studies, even by only considering 'untargeted' attacks. The literature has achieved undeniably valuable results with remarkable precision in simulating the archetype of SE attacks: a phishing email with a malicious link. However, there are many more scenarios that are worth investigating deeply, which are largely ignored by the literature (cf. Section 4.4.3). Moreover, several simulated phishing campaigns and classification tasks do not reflect the current threat landscape (cf. Section 4.4.5). Future experiments should account for *multi-step* attacks that go beyond clicks only, such as submissions of credentials with multi-factor authentication (e.g., MFA bombing [29]) or opening attachment-like artifacts. Lab experiments should accurately reproduce

the complete attack process, e.g. click then submit, and allow for interactive interfaces (such as hovering on links, reactive forms, etc.) over static screenshots of emails and websites. Further, modern-day attacks feature a diversification of utilized media (cf. Section 4.4.3) and of the modality of their employment (cf. Section 4.4.5). Therefore, the need to precisely simulate complex attack scenarios extends not only to multi-step, but also to *multi-modal* simulations where attack interactions may cross multiple media, applications, and devices; for example, combinations of instant messaging and websites [137], social networks and email [15] or QR codes [78, 354]. In addition, the gap between real attacks and studies in empirical SE stems from the limited scoping of experiments to specific domains. For example, the majority of experiments are conducted with participants drawn from university pools (cf. Section 4.4.2), while companies and institutions belonging to other domains, such as governmental or industrial, are overall under-represented, albeit increasingly at risk of generic, spear-phishing, and tailored campaigns [39]. Yet, studies already observe that the effects of attacks can significantly vary *across* organizations operating in different domains and, at the same time, across different roles in an organization [61, 132, 196]. We underline that assessing the state of target susceptibility in richer multi-step scenarios across different domains and how new media can be weaponized is a necessary path toward filling the gap between real and simulated SE attacks.

The SE attack surface is vast

While the gaps between real and simulated attacks mainly concern *how* simulations are carried out, the gaps in the coverage of the SE attack surface regarding *which* attack dimensions, and relative combinations, have been investigated in the literature. Our study points out that the attack surface available to the attackers is vaster than what the experiments covered thus far, and its exploitation by real attackers is growing wider. As shown in Section 4.4.5, work-related and setting-related target parameters are seldom investigated in the literature, perhaps because they are not easy to isolate and control, particularly in *in vivo* experimental settings. Nonetheless, email load, timing, and, most of all, the relevant dimensions of a social or communicative situation are shown to significantly affect the susceptibility of the targets [132, 320], and represent good candidates for experimentation in future research. The most uncharted segments of the SE attack surface are the patterns and nuances of human cognitive systems and, as highlighted in Section 4.4.5, the effects of priming and specificity of percepts are still unclear. Similarly, studies investigating attention suggest that it can play a decisive role in the outcome of an SE attack (cf. Section 4.4.5), although possible vulnerabilities related to attention remain, at the moment, largely unexplored. Heuristics represent another important avenue of exploitation of the human attack surface and have been moderately studied in the literature, but almost exclusively as indirect effects of stimuli attributes (cf. Section 4.4.5). The dynamics regulating *Heuristics* and *Anomalies*, and their interplay, require additional research particularly to evaluate the effect of different pretexts, attack/target parameters, and multi-stage attacks on attack success (and defense effectiveness). For example, whereas the literature suggests systematic processing is beneficial to thwart phishing attacks [145], it remains unclear under which circumstance it is triggered during processing. Similarly, the power of 'anchoring' effects such as cognitive biases is unclear, and specific methods to alleviate their effect have not been explored at the moment.

The exploitable attack surface, thus, appears to be much larger than the current coverage provided by the state-of-the-art. Addressing these gaps would allow an unprecedented understanding of the human attack surface that enables SE attacks in the first place and support the design of novel prevention techniques. To this end, we encourage further experimentation covering uncommon attack scenarios of higher risk in the current threat landscape. Moreover, we recommend expanding previous approaches to new or adapted techniques from cognitive science and social psychology (e.g., [207, 254]) with the aim of capturing and analyzing the target's contextual and cognitive factors.

Studies are focused on a few experimental setups only

We find that the literature tends to employ certain experimental methods with specific populations. Section 4.4.2 suggests that subjects from the general public are often associated with lab experiments and non-university staff with, predominantly, field experiments. This may depend on the limitations of recruiting procedures (e.g., for ethical reasons) or the need for a controlled environment. Moreover, the reviewed literature tends to employ population-level targetization almost exclusively with field experiments, while generic-level targetization is addressed by other types of studies, mainly laboratory experiments (cf. Section 4.4.4). This can make the obtained results of limited explanatory power. In addition, the common methods to carry out SE experiments may not be suitable to test hypotheses involving a variety of cognitive factors (cf. Section 4.4.5). This saturation of what can be measured or tested does not help fill the gaps between real and simulated attacks and to cover uncharted segments of the SE attack surface. Some of these limitations may be mitigated for experiments in organization settings, where the investigators may have access to fine-grained data to control (e.g., seniority or operational setting, cf. Section 4.4.5) or measuring confounding variables (e.g., contextual factors specific to that organization). Field experiments with the general public may be unattainable for ethical reasons. However, the research may gain similar insights with observational studies, perhaps in collaboration with service providers as in [10, 391]. As some cognitive features cannot be measured quantitatively and potential confounding factors cannot be fully controlled for, qualitative insights (especially as enabled by surveys and interviews) can shed further light on contextual factors as well as to qualify the effects beyond merely the metric of choice. Therefore, we advocate for the inclusion of such instruments as part of the 'standard' phishing experiment setup.

Lack of common reference for targetization

We find that the literature is inconsistently referring to targeted attacks while employing only general or population-level targetization. For example, the experiments in [45] and [244] self-report the use of individual and population-level targetization respectively, while these studies are classified as generic-level targetization according to the criteria of Section 4.3.3. More in general, the adaptation of stimuli to the participants depends on the target's environmental and contextual factors which, in turn, are difficult to reproduce across repeated measurements [132], as also thoroughly discussed in Section 4.4.5. This can draw confusion on the state of research with respect to some types of SE attacks and the used terminology, such as phishing or spear-phishing, where the reported results are, if not contrasting,

inconsistent. For example, many studies tested spear-phishing, "social phishing" or other variants of targeted stimuli yielding a significantly higher success rate vs. un-targeted control groups [125, 137, 161]. However, there were also reports of targetization not yielding increased attack success rates where, for example, users were more susceptible to emails with links to external servers than they were to email with links to internal servers [48], and where attempts to individualize adaptations (e.g., saluting the recipient by name [164] or congruently to expectations of participants [250]) were no more successful than generic emails [320]. The lack of approaches for consistently gauging the level of the inherent targetization of the employed stimuli against a common reference scale, or at least a common definition of targetization, makes a coherent interpretation and comparison of these conflicting findings impractical. Therefore, we identify the need for upcoming SE frameworks to systematically enable the measuring of sophistication (and thus targetization) levels of SE attacks. A first step in this direction is provided by the framework in Chapter 3, which evaluates the parameters assumed by the attacker vs. those of the subjects, thus providing a consistent accounting of the adaptation degree between stimuli, experiment subjects and context.

Inconsistent constructs of experimental outcomes with respect to the current threat landscape

Our analysis shows an overall inconsistency in how the extant literature defines a successful SE attack. The experimental constructs devised to measure the success rate vary, from study to study, between clicking a link, opening an attachment, visiting a webpage, submitting credentials, answering to an email, etc. However, each of these constructs arguably measures different degrees of attack success and, conversely, leads to conflicting findings. This particularly concerns field experiments where the attack success is approximated with clicks on links, which do not necessarily lead to a security impact, as discussed in Section 4.4.5. Clicks or wrong classification of stimuli may result in a successful outcome for an attacker, and studies employing such measurements provide valuable results [270, 314]. However, clicking on links equals attack success, or low classification accuracy equals high susceptibility, have become a common assumption that can lead to wrong or imprecise conclusions. Especially in the field of Information Systems or other disciplines that also investigate SE attacks (e.g., medicine [166], behavior sciences [154]) click rates are often adopted as the de-facto measure of attack success. The inconsistency of attack outcome measurements can bear important consequences on how research results are applied in practice. Indeed, organizations might use the results of simulations, i.e. click rates, to draw conclusions on their information security posture and to guide company policies, such as budget allocation or reprimand of employees [187, 289]. However, when not applied carefully (e.g., without considering potential confounding factors, such as users' motivations, competing tasks or misinterpretation of training material [93]), this can lead to adverse side effects, such as a false sense of security, making employees even more vulnerable to phishing [132, 196].

Relevant factors often not controlled for

As we have seen in Section 4.4.5, some factors are especially difficult to control or measure, namely setting parameters as well as cognitive features such as *Perception*, *Attention* and *Elab-*

oration. For example, keeping track of a large number of targets' primary goals or concurrent events in a given time frame would require an enormous monitoring effort, or measuring the state of cognitive features may be too invasive and infeasible with a large number of participants [26]. However, we argue that there are numerous opportunities to face such challenges by looking into the fields of cognitive science and (social) psychology. Supplementary techniques from such fields can provide valuable methodological insights to investigate, for example, various effects of different contextual variables in SE experiments [226], the role of priming [90, 374] and memory [312] in subjects' perception or how their elaboration can be influenced by attention [207], heuristics [44, 134], and anomalies [153]. Such an interdisciplinary approach for solving problems in information security is well exemplified by the results obtained in usable security studies where concepts and techniques, such as behavior change theories [231] and digital 'nudges' [333], have been successfully employed in experiments to evaluate, for example, warning [208, 249] or training efficacy [192, 318], as also discussed in [121, 276].

Considerations for practice

Our findings point in the general direction of a gap in measurements of human attack surface over several dimensions. For example, whereas training and awareness programs at organizations mostly focus on emails [317, 336], consideration should also be given to other communication channels used at the organization, such as SNSs or messaging apps. Further, our findings suggest that specific outcome measurements may not always be representative of the actual threat. For example, carrying out embedded phishing exercises to assess the state of target susceptibility with click rates may not necessarily represent a realistic compromise scenario even if these are used 'interchangeably' in the literature with actual credential submissions as a measure of 'attack success'. Similarly, any specific countermeasures already in place should be accounted for in the experimental setting; for example, measuring credential submissions may provide an unrealistic picture of the overall threat if Multi-Factor Authentication protocols are in place. Similarly, accounting for the confounding factors that lead to a given attack outcome and the embedded training side-effects, can strengthen the understanding of why users fall for the attack and assist the implementation of appropriate security controls or processes. For example, the success rate of an instance of embedded phishing exercise can be heavily conditioned by the type of pretext and user context in a given moment [66, 132]. Understanding these elements can assist the design of anti-phishing training and awareness programs tailored to the specifics of the organization's domain or even of employee characteristics [222]. Similarly, embedded phishing training may unintentionally train employees to not report phishing emails after interacting with them [93], whereas reporting phishing to IT is a desired behavior. Practitioners may want to complement embedded training with measurements (e.g., briefly after the phishing email) of user context to understand what conditioned a user's behavior and, therefore, to encourage or discourage a certain action, such as reporting to IT vs. asking for confirmation to the sender [93].

4.5.1. Threats to validity

Internal threats. The cognitive framework adopted for our analysis (see Chapter 3) was distilled from mainstream theories and models in cognitive science. Being this a diverse field with sometimes inconsistent usage of concepts and with ongoing debates, the perspectives in cognitive science may vary with respect to different debates, which are far away from closed. Nevertheless, the framework abides by the most shared views in the field of cognitive science and lends itself to a generic enough application to SE to avoid such risks, akin to what is done in previous work [238], and as discussed in Appendix B.

External threats. The search query used on the Scopus database to retrieve the body of empirical research could have missed some relevant papers. We mitigated such limitations by performing a reverse snowballing till saturation was reached, and by running more specific queries on Scopus which showed no substantial difference in results. An additional limitation may be the query restriction to the Computer Science subject area. To verify this, we performed additional manual checks (including further snowballing) on the results from the same query without restriction and found that the limited set of additional papers resulting from the procedure did not significantly change the results presented in this paper. Further, we encountered two main fields that contained the bulk of the reviewed papers: IT Security and Information Systems. The description detail, scoping, and comprehensiveness of publications in such fields may vary and, thus, condition the applicability of inclusion and analysis criteria. However, this is a reflection of the actual state of affairs in SE research and we have no reason to believe that our method missed other research fields (e.g., Human-Computer Interaction, Decision Support Systems, Computer-Mediated Communications) that regard (empirical) SE with particular interest as the previous two. This suggests that the collected sample well represents the current state of the art of empirical SE research on cognition.

Construct threats. Some works investigated variables related to Elaboration, Heuristics, and Anomaly with multiple other features in the same hypothesis. This can result in associations that might appear counter-intuitive, e.g. cognitive biases are related to pretext and look&feel in Fig. 4.14. We argue that this still reflects the original intentions of such papers and does not affect qualitatively the results of our review. Also, there could be bias and subjectivity in the extraction and grouping of some variables, e.g., the clear-cut classification of study types, the adaptation levels of stimuli, or even features of cognitive systems. The authors iteratively discussed and confronted the ambiguous situations until a consensus was reached (more detail in Appendix B). More in general, this is a common problem in similar articles, where the absence of an established framework for classifying SE experiments poses such limitations [320]. To this end, we distilled our criteria from a cognitive framework specifically designed for the analysis of SE attacks, akin to the previous work discussed in Chapter 3.

4.6. Conclusion

This study provides a systematic review of the state of empirical SE research on cognition with the goal of advancing the body of knowledge in the SE domain by identifying and characterizing the open gaps between the features of human cognitive processes and empirical research in SE. To this end, we systematically analyzed 169 articles from the wide and multidis-

ciplinary landscape of empirical SE research along the dimensions of experiment design and human cognition. Our study reveals that experiments only partially reproduce real attacks and that the exploitable SE attack surface appears much larger than the coverage provided by the current body of empirical research. Factors such as targets' context and cognitive processes are often ignored or not explicitly considered in experimental designs. Similarly, the effects of different pretexts and varied targetization levels are overall marginally investigated.

By relating the findings of the analysis with the dynamics of real attacks and extant SE research, we identified open problems, relevant insights and promising directions for future work. This provides an answer to RQII. Stemming from the considerations in Chapters 3 and 4. In the next part of this thesis (Chapters 5, 6 and 7) we address the problem of tailored phishing attacks and potential phishing mitigation strategies (RQIII).

II

Tailored phishing and potential counter-strategies

5

Tailored phishing attacks

As the email remains the main SE attack vector, phishing attacks keep evolving into more sophisticated and targeted variants. We have seen in Chapter 2 that tailored phishing represents a lightweight, scalable variant of spear-phishing, which can lead to a high impact with relatively low effort from the attacker. In this scenario, attackers tailor their phishing emails to increase attack credibility based on the information about the intended victims at the scale of entire organizations. Among the gaps in empirical SE research identified in Chapter 4, it remains unclear to what extent known tailored phishing techniques improve attack success rates, and their interaction effects with well-known persuasion techniques, as well as targeted populations. In this chapter, we report a field experiment targeting 747 participants employed in two organizations (a university and a large international consultancy company) to evaluate the interaction between phishing persuasion techniques and the attack success rate in a highly-tailored setting (see RQIII). For this purpose, we exploit well-established *user notification* methods to enhance the delivery of persuasion techniques (e.g., Authority, Scarcity, etc.), and evaluate how such techniques affect the phishing success rate across industrial and academic domains. We find that the effect of 'traditional' attack techniques, such as those relying on cognitive vulnerabilities, is widely mitigated in highly-tailored phishing settings, suggesting that current user training and detection techniques may be off-target for more sophisticated attacks. However, we find that the *means* by which the attack is delivered to the victim matter, and can greatly (up to three times) boost the effect of the base attack.

This chapter is originally published as P. Burda, T. Chotza, L. Allodi, and N. Zannone, "Testing the Effectiveness of Tailored Phishing Techniques in Industry and Academia: A Field Experiment", In the *15th International Conference on Availability, Reliabilityand Security* (ARES), ACM, 2020, pp. 1–10

5.1. Introduction

We have seen in Chapters 1 and 2, that phishing is becoming the most prevalent source of compromise for most organizations [350, 352]. Novel social-engineering models where attackers move *laterally* within an organization to increase the effectiveness of their attacks [148], and sophisticated, multi-stage attacks targeting representatives of social minorities in China [47], as well as at scale [15], have been reported in the literature.

A key element of this trend is that attackers gather targeted information about their victims, and use it to build *tailored* phishing attacks targeting that victim (or group of victims) specifically. Among the gaps identified in Chapter 4, the techniques attackers can employ to best exploit this information are still unexplored. The effectiveness of well-known cognitive effects in well-tailored attacks is unclear [347, 369, 379], and the scientific evidence is still inconsistent, for example, across application domains [82, 320] or job roles [132]. To answer these questions we must develop specialized experiments measuring the effectiveness of sophisticated phishing techniques on real users.

5

In this chapter, we provide the first insights on the relation between tailored-phishing, cognitive attacks, their delivery methods, and the organizational settings in which the victims operate. We perform two simulated phishing campaigns against three departments of a mid-sized European University (`UNI`), and a division of a leading, international company operating in the consultancy sector (`IND`). We derive different attack delivery techniques from the *user notifications* literature [391], and identify the relation between notification techniques and cognitive attacks commonly employed in phishing [347, 369]. We ran our experiment in both `UNI` and `IND` in June 2019, targeting $n = 747$ employees spanning `Senior`, `Junior` and `Support` roles within the respective organizations, and measure the relative efficacy of the adopted 'enhanced' phishing techniques.

Following on RQIII, we structure this chapter along four specific research questions:

- RQ1 Does professional role have an impact on success rate of tailored phishing attacks?
- RQ2 Are cognitive vulnerabilities effective in tailored phishing?
- RQ3 Do notification methods to deliver cognitive vulnerabilities increase the success rate of a tailored phishing attack?
- RQ4 Which characteristics of the campaign and victims affect the velocity of a tailored phishing attack?

Our contribution can be summarized in three points:

1) We identify *notification methods* (the *way* a persuasion technique is implemented) that can be used by attackers to enhance the effectiveness of phishing attacks, and their relation to well-known persuasion techniques adopted by attackers. We find that 'baseline' cognitive attacks are not effective in tailored phishing scenarios, whereas the adoption of notification methods can boost attack success (up to three times).

2) We provide insights on the effect stability of different well-established as well as novel phishing techniques (i.e., combinations of persuasion techniques and notification methods) across organizations operating in different domains, and employee expertise. We show that

the effectiveness of some attack technique can significantly vary across organizations operating in different domains.

3) We measure and report the *velocity* at which users fall for our attack, how that diverges by user category and attack type, and provide insights for effective attack response practices.

NB: The phishing experiment described herein has been characterized in Chapter 3 according to the cognitive framework (introduced in Chapter 3) where the mapping of experiment treatments to the features of cognition have been discussed in detail. For example, the effects of the job role and target domain is hypothesized to influence the participants' ability to detect the fraudulent nature of the phishing pretext (e.g., `Junior` vs. `Senior` roles). The interested reader may refer to Chapter 3 for a detailed reasoning on which features of the current experiment might affect the cognitive processes of the participants.

The chapter is structured as follows: Section 5.2 provides background on cognitive vulnerabilities and user notification techniques. Sections 5.3 and 5.4 detail the adopted methodology and its implementation respectively. Results are presented in Section 5.5, whereas Section 5.6 discusses our findings.

5.2. Background and Related work

In Chapter 4, we have observed that various empirical studies have been conducted to assess the efficacy of phishing campaigns, e.g., [11, 52, 161, 164, 256]. Such studies can be roughly classified in three types: observational, laboratory experiments and field experiments (with surveys and interviews as additional data collection methods). The efficacy of a phishing campaign is usually evaluated with respect to three main factors, namely attributes of the phishing email, victim characteristics and context, and by measuring some form of participants' behavior, as thoroughly discussed in Chapter 3. In this work, we perform a field experiment, focusing on the attributes of the phishing email and victims' characteristics, and measure the attempts of participants to login on a bogus web page (i.e., the payload of the email).

5.2.1. Cognitive vulnerabilities

Cialdini's *principles of persuasion* [73] are commonly employed in the literature as predictors for phishing success. These principles define six 'cognitive vulnerabilities' that can be exploited to influence individual decision-making: `Authority`, `Liking`, `Scarcity`, `Consistency`, `Social proof` and `Reciprocity`. An overview of the principles of persuasion is presented in Table 5.1. These principles are employed across the spectrum of human communication activities. For example, in marketing `Scarcity` can be exploited by limiting the duration of sales, generating fear of missing out and pushing clients to purchase more than they would otherwise do.

Cognitive vulnerabilities are often exploited in phishing campaigns, where `Liking` and `Authority` appear to be the most used [371]. However, their effectiveness has been shown to vary significantly across studies. For example, `Liking` and `Social proof` were reported to have a positive effect on phishing success [379]. `Authority` was highly effective according to [63, 155, 371] and `Scarcity` to [371]. Other studies find 'reversed'

Table 5.1: Cognitive Vulnerabilities

Principle	Description
`Authority`	Tendency to obey people in authoritative positions, driven by the possibility of punishment for not complying with authoritative requests.
`Liking`	Tendency for saying 'yes' to requests of people they know and like. People are 'programmed' to like others who like them back and who are similar to them.
`Scarcity`	Tendency to assign more value to items and opportunities when their availability is limited.
`Consistency`	Tendency to behave in a way consistent with past decisions and behavior. After committing to a certain view, company or product, people will act in accordance with those commitments.
`Social proof`	Propensity to label behavior as correct to the degree that others performing it.
`Reciprocity`	Tendency to repay, in kind, others for a received favor.

effects, whereby the presence of certain cognitive vulnerabilities *reduces* the probability of success of the attack. For example, van der Heijden et al. [347] find a negative correlation between `Reciprocity` and the number of users falling for the attack in the banking domain. Other studies purport similar negative effects for `Authority` [379] and `Social proof` [63]. Other studies reported no significant effect in either directions for `Consistency` and `Reciprocity` [379].

These contrasting results provide mixed evidence on the relative effectiveness and employment of cognitive vulnerabilities in phishing attacks. It appears that *context* matters [132], suggesting that the way in which these cognitive attacks are delivered to victims has an impact on their effectiveness [114]. Oliveira et al. [256] conducted a field experiment to test the relative effect of cognitive vulnerabilities, age and life domains (private context and different pretexts) on spear phishing susceptibility. However, no studies to date evaluate how different implementations of the same cognitive attack vary the effect and, in particular, how 'stable' such an effect can be expected to be across varying contexts (e.g., organizations).

Similarly to our work, Butavicus et al. [63] have studied the effectiveness of cognitive vulnerabilities in general phishing vs. spear-phishing. Whereas [63] adopts a survey-based lab experiment involving 121 undergraduate students, we perform an ecologically diverse field experiment to evaluate how the presence and delivery of cognitive vulnerabilities affect tailored phishing success.

5.2.2. Notification methods

Part of the reason why different techniques for the *delivery* of cognitive attacks in phishing have not been investigated yet is the lack of a framework to consistently differentiate attack features (cf. Chapter 3). On the other hand, the problem of effectively conveying a cognitive attack is not dissimilar from that of effectively *notifying* a risk or security measure: in both cases the sender needs to find the most effective way of obtaining compliance from the user [391].

Table 5.2: Notification Methods

Principle	Description
Contact information	Contact information includes a name, an email address, a business address, phone numbers and more, and is typically added at the bottom of an email. Including contact information in the email increases credibility to the notification [379].
Personalization	Personalizing the email content to the identity of the recipient, her location, and preferred language increases the feeling that the notification is custom-made for the recipient [79].
Subject Line	A semantically meaningful and captivating subject line contributes to capture the attention of the recipient and thus help increasing email opening rates [178].
Detailed information	Providing detailed information about the addressed topic gives a sense of proactivity [349].
Time	Notifications should be sent when recipients are most open to receive them [382].

Several studies investigated the factors affecting users' reply behavior in organizational settings [68, 205, 349, 382]. Table 5.2 provides an overview of the relevant notification methods identified in those studies. For example, sending *time* is known to have a significant impact on the rate of notification responses [382]. Similarly, Li et al. [205] and Vasek and Moore [349] show that *detailed information* increases the remediation rate of security issues communicated to end-users, when compared to the effectiveness of more terse notifications.

Other aspects that affect notification effectiveness are found in practices and guidelines for email and push notifications. Some studies show that basic *personalization* of notifications results can considerable increase email opening rates [79]. Similarly, including *contact information* in the email is considered to be part of email etiquette, and may contribute to increase the credibility and relevance of the message to the receiver [21, 214]. However, the authority of the signatory does not always have an effect on user compliance [68]. By contrast, clear headings in the *subject line* can (but not always [178]) significantly affect user behavior [168, 218].

In this work, we evaluate whether *notification methods* can be used by attackers to enhance the effect of *cognitive vulnerabilities*, and whether that effect varies depending on the context. For brevity, hereafter we will refer to the combination of notification techniques and cognitive vulnerabilities as *persuasion techniques*.

5.3. Methodology

To determine which persuasion techniques have an impact on phishing success, we performed a field experiment on 747 subjects. Selected subjects are randomly assigned to an experiment condition, namely a persuasion technique composed of a cognitive vulnerability and a notification method. The adopted phishing pretext leads users to login to a web page resembling their organization's login portal. Phishing success is measured in terms of rates of users that provide their credentials (cf. Sec. 5.3.4 for ethical aspects).

Table 5.3: Employee categories at UNI and IND.

Category	Org.	Job function	no.
Junior	UNI	PhD students, Postdocs	690
	IND	Non-Managerial staff, interns	723
Senior	UNI	Faculty members (Assistant, Associate, Full Prof.)	420
	IND	Managerial staff	480
Support	UNI	Research support, secretaries, admin staff	210
	IND	Technical specialists, secretaries, mgmt. assistants	180

5.3.1. Experimental approach

Organizational settings

To assure the ecological validity of our experiments, we launch two phishing campaigns in two diverse settings: a mid-sized university (UNI) and a European branch of large consultancy company (IND). In particular, we target three departments of UNI (total of 1320 employees),[1] and one division of IND (total of 1383 employees). The contrast between IND and UNI allows us to evaluate treatment effects across two different organizational settings. IND runs regular phishing campaigns as part of security awareness training for its employees. By contrast, UNI does not employ any form of security awareness training for its staff.

Sampling

Employees at both organizations have different levels of professional expertise and job functions. For both UNI and IND, consistently with dimensions identified in previous studies [52, 256], we group employees into three categories according on their seniority within the respective institution: Junior, Senior and Support. Table 5.3 provides a mapping of job functions to the respective categories for UNI and IND. We limit our sample to 30% of employees in each of the Junior, Senior and Support categories; this minimizes 'word of mouth' effects that may affect experiment outcome.

Persuasion techniques

We adopt a selection of persuasion techniques in the cognitive and notification domains (ref. Sec. 5.2.1 and 5.2.2) as the treatments for our experiment. Among the cognitive vulnerabilities identified in [73], we adopt the Authority, Liking, Scarcity and Consistency principles. We exclude Social proof and Reciprocity, as these relate to social pressure (i.e., behaving consistently to peers) rather than to professional customs (i.e., related to professional duties and obligations).

To evaluate how cognitive effects are mediated by the adopted notification method, we employ a second set of treatments modifying the implementation of the cognitive attacks. Ta-

[1]The selected departments are Chemical Engineering and Chemistry, Industrial Design, and Mathematics & Computer Science. These were chosen to cover a wide spectrum of computer systems knowledge, with Chemical Engineering on the lower end and Computer Science on the higher end of the spectrum.

Table 5.4: Association between selected cognitive vulnerabilities and notification methods.

Cognitive vuln.	Notification method	Association rationale
`Authority`	*Contact information*	The inclusion of *contact information* in an email conveys a sense of formality that the recipient might perceive as `Authority` and thus might make them feel more inclined to comply to the request.
`Liking`	*Personalization*	Notifications should be tailored to each user to create a baseline of one-to-one familiarity. Based on the insights of previous studies [52], we use the recipient's first name in the message to create such sense of familiarity and, thus, to induce a perception of `Liking`.
`Scarcity`	*Subject line*	The 'feeling of urgency' exploited by `Scarcity` can be immediately conveyed through the subject line, which may provide an additional effect on the victim opening and acting on the email. This feeling can be generated by including words implying time sensitivity and triggering an immediate call to action in the subject lines, like "urgent", "important", "alert", etc.
`Consistency`	*Subject line*	The content of a subject line can also be exploited to describe follow-up actions to previous interactions, thus conveying a sense of `Consistency` with previous decisions and commitments. Common examples are `Re:` formats employed in subject lines of phishing attacks.

ble 5.4 reports mapping and rationale for the association of the adopted cognitive vulnerabilities with the relevant notification methods. These are chosen to match the functional purpose of the respective cognitive vulnerability. For example, *contact information* can exercise `Authority` on a victim, whereas *subject line* is ill-suited to convey primarily authoritative message.

5.3.2. Experiment conditions

Based on the persuasion techniques above, we devise nine experiment conditions, summarized in Table 5.5. To synthetically represent them, we adopt the following notation: each condition is assigned an identifier of the form R^N_C, with $N \in \{D, A\}$ reflecting a default (D) or advanced (A) notification method, and $C \in \{Au, Lk, Sc, Cn\}$ indicating the presence of a specific cognitive attack: `Authority` (Au), `Liking` (Lk), `Scarcity` (Sc), `Consistency` (Cn). The baseline condition with no cognitive exploits is denoted as R^D.

5.3.3. Evaluation Criteria

Criteria for RQ1-RQ3

We measure the success of a phishing campaign in terms of the rate of subjects that submitted their credentials. Specifically, the probability of success $p^k_{i,j}$ per user group i (`Junior`, `Senior`, `Support`) at an organization k (`UNI`, `IND`) and experiment condition j (R^D, R^D_{Au},..., R^D_{Cn}, R^A_{Cn}) is computed as the number of successful outcomes (i.e., subjects that submitted

Table 5.5: Experiment conditions

Notification Method	None	Authority	Liking	Scarcity	Consistency
None	R^D	R^D_{Au}	R^D_{Lk}	R^D_{Sc}	R^D_{Cn}
Contact info.	—	R^A_{Au}	—	—	—
Personalization	—	—	R^A_{Lk}	—	—
Subject line	—	—	—	R^A_{Sc}	R^A_{Cn}

their credentials) over the total possible outcomes (i.e., sent emails):

$$p^k_{i,j} = \frac{(\#\ submissions)^k_{i,j}}{(\#\ sent\ emails)^k_{i,j}}$$

To answer research questions RQ1 to RQ3, we employ *odds ratios* (*OR*), which quantify the effect size of a treatment by measuring the relative change in the probability of success between the treatment group and the baseline:

$$OR = \frac{p_I/(1-p_I)}{p_B/(1-p_B)}$$

where p_I and p_B are respectively the probability of success for the group receiving the treatment and for the group receiving the default email. $OR = 1$ means that the probability is the same regardless of the presence or absence of the treatment (i.e., the treatment has no effect); a ratio greater (lower) than 1 indicates that the treatment has a positive (negative) effect on the success of the campaign compared to the baseline.

For RQ1, we compute the probability of success per each user category i at an organization k as:

$$p^k_i = \frac{\sum_{j\in\{R^D,\ldots,R^A_{Cn}\}}(\#\ submissions)^k_{i,j}}{\sum_{j\in\{R^D,\ldots,R^A_{Cn}\}}(\#\ sent\ emails)^k_{i,j}}$$

For RQ2 and RQ3, we compute the probability of success per each experiment condition j at an organization k as:

$$p^k_j = \frac{\sum_{i\in\{\texttt{Junior},\texttt{Senior},\texttt{Support}\}}(\#\ submissions)^k_{i,j}}{\sum_{i\in\{\texttt{Junior},\texttt{Senior},\texttt{Support}\}}(\#\ sent\ emails)^k_{i,j}}$$

Statistical significance. To report the statistical significance of our results we compute 95% confidence intervals (CI) for the odds ratios. A result is considered statistically significant when the calculated confidence interval is either completely below (*negative* effect) or completely above (*positive* effect) the unity. To *qualitatively* evaluate marginally significant results, we additionally compute 80% confidence intervals (we do not report these for the lack of space). We evaluate our results both *within* and *across* the participating organizations: the former indicates differences across experiment conditions within the same organization (CI completely above or below unity); the latter compares differences in outcomes for the same

experiment condition between the two organizations (*OR* point estimate for one organization is outside of CI of the other organization, and *vice versa*). To graphically differentiate reporting of highly significant and marginally significant results *within* and *across* organizations, we adopt the following notation:

	Strongly sig.	Marginally sig.
Across	green cells	yellow cells
Within	**bold, black text**	**bold, blue text**

The table below provides an example of result interpretation (mock experimental outcomes). For simplicity, we only report an example for *strongly* significant results:

	OR [lower, upper]	
Cond.	$\texttt{ORG}_1$	$\texttt{ORG}_2$
A	**0.6 [0.2,0.7]**	0.9 [0.8,1.2]
B	**1.3 [1.1,1.4]**	**1.5 [1.1,1.8]**

Within condition *A*, $\texttt{ORG}_1$ shows a *significant, negative* effect w.r.t. the baseline: the *OR* point estimate (0.6) indicates that chances of phishing success in experiment condition *A* are 40% *lower* than in the baseline condition for $\texttt{ORG}_1$, and as the CI does not cross the unity the result is statistically significant. By contrast, the effect for $\texttt{ORG}_2$ is not significant. Comparing the effect across organizations, the point estimate for $\texttt{ORG}_1$ (0.6) is *outside* of the CI of the estimate for $\texttt{ORG}_2$; the opposite is also true. This indicates that experiment condition *A* has also a *significantly different* effect across the two organizations.[2] As the point estimate of the former is *lower* than the point estimate of the latter, this indicates a negative effect in $\texttt{ORG}_1$ when compared to $\texttt{ORG}_2$ under condition *A*. By contrast, under condition *B* estimates for both $\texttt{ORG}_1$ and $\texttt{ORG}_2$ are significant *within* the respective organizations, but no significant difference emerges *across* organizations as the *OR* estimate for $\texttt{ORG}_1$ (1.3) falls within the CI of the effect of *B* in $\texttt{ORG}_2$ ([1.1, 1.8]).

Criteria for RQ4. Research question RQ4 aims to determine which characteristics of the campaign and victims has an effect on the velocity of the attack. To this end, we compute how the success rate progresses over time.

5.3.4. Ethical considerations

The experiment was approved by the UNI's Ethical Committee and by management at IND. We followed best practices concerning consent waving and user debriefing [293]. 'Submitted' user credentials as well as the association between user identities and their real names were neither transmitted nor saved by the system.

[2] In the table we denote that a condition exhibits significant effects *both* within an organization ($\texttt{ORG}_1$ in the example) and across organizations by using both the bold font and the green cell.

5.4. Experiment design

This section describes the design and implementation of the experiment. First, we present the scenario of the phishing campaign and then we discuss the concrete implementation and data collection.

5.4.1. Experiment preparation

This section describes the scenario and artifacts (phishing emails, website) developed for our phishing campaigns.

Scenario selection and attack prototype

We design two phishing campaigns in close collaboration with the security and privacy teams of UNI and IND to assure their credibility. Both campaigns are built around the same *pretext*, namely the need to register holiday hours within the organization's portal for administrative purposes. We adopt the following phishing email prototype:


```
From: info@{domain-name}
Subject: Your holiday hours
Dear Colleague,
To facilitate the planning of activities for the period September to De-
cember, we invite you to provide a rough estimate of the holiday hours
you are currently planning to take until the end of this calendar year.
Please provide this information by following this link: {domain-
name/path}
Thank you,
{signature}
```

This prototype is used for both UNI and IND. The values in {curly braces} were customized for each organization using a signature and a domain name (for both the link and From field) that resemble the ones typically used in those organizations; no spoofing techniques was used. Values were defined jointly with the security and privacy teams at UNI and IND. All experimental treatments are applied to this baseline email.

Treatments

We defined a treatment for each experiment condition (ref. Sec. 5.3.2). A treatment is a modifier to the baseline email. We consider two sets of treatments: one set in which cognitive vulnerabilities are added to the body of the baseline email prototype, and one set adopting a specific notification method to deliver the cognitive vulnerability. An overview of the administered treatments is presented in Table 5.6.[3] For example, the treatment for Scarcity is implemented by extending the body of R^D with a deadline for the submission of holiday hours ("by the end of this week") (R^D_{Sc} in Table 5.6). The notification method treatment is then applied to R^D_{Sc} by including "Action Required" in the subject line (R^A_{Sc} in Table 5.6). This allows us to evaluate the relative increase in effectiveness of *adding* a

[3]The emails are available at https://github.com/paolokoelio/testing-tailored-phishing

Table 5.6: Principles used in the campaign and respective implementations.

The treatments in which the cognitive attacks are delivered in the body of the prototype email are presented on the left side. The treatments in which the cognitive attacks are delivered through notification methods are presented on the right side. Changes from the respective baseline email are highlighted in **bold**. The meeting in R_{Cn} is fictitious.

Cognitive Attack			Notification Method		
ID	**Principle**	**Implementation**	**ID**	**Principle**	**Implementation**
R^D_{Au}	Authority	"[...] to December, **the {ORG} Support team, on behalf of the Executive Board, requires** you to provide a rough estimate [...]"	R^A_{Au}	*Contact Information*	**From: secretariaat.executive@{domain-name}** Subject: Your holiday hours "Dear Colleague, To facilitate the planning [...] **All inquiries must be directed to: secretariaat.executive@{domain-name}** Thank you, **The {ORG} Executive Board Secretariaat** **Email: secretariaat.executive@{domain-name}** **Address: {ORG's street, ZIP, City}**"
R^D_{Lk}	Liking	"[...] to December, **and to take into consideration your future plans**, we invite you to provide a rough estimate [...]"	R^A_{Lk}	*Personalization*	"Dear {**FirstName**}, To facilitate the planning [...]"
R^D_{Sc}	Scarcity	"[...] to December, we invite you to provide **by the end of this week** a rough estimate [...]"	R^A_{Sc}	*Subject Line*	Subject: Your holiday hours **- Action Required**
R^D_{Cn}	Consistency	"Dear Colleague, **This is a follow up to the {Department} Employee Meeting held past February.** To facilitate the planning [...]"	R^A_{Cn}	*Subject Line*	Subject: Your holiday hours **- Follow up**

specific notification method to a predefined cognitive attack. The implementation of the selected cognitive attacks is consistent with wording used in past experiments and field studies reported in the literature [15, 148, 347, 379].

Phishing website

We crafted a phishing website for each organization consisting of a login page mirroring that of the corresponding organization. The webpages are instrumented to log the loading of the phishing website, and the credential submission. Upon submitting their credentials, users are redirected to a debriefing web page. The debriefing page informs victims about the experiment (i.e., purpose, authors, which data have been collected and how it will be used, and experiment authorization). Author contact information was also provided for any additional communication and to waive consent to data usage.

5.4.2. Implementation and data collection

The sampling method described in Sec. 5.3.1 results in 396 individuals sampled from UNI and 415 from IND; however, a group of employees in IND ($n = 64$) proved to be 'immune' (by design) to the chosen pretext (see Sec. 5.6 for a discussion), and we therefore exclude them from the analysis. This leaves us with 396 and 351 subjects in UNI and IND respectively, for a total of 747 subjects. Each sampled subject was randomly assigned to an experiment condition and received exactly one email in the context of this experiment. Accordingly, 83 subjects were assigned to each condition (44 at UNI and 39 at IND). Sub-

jects were neither informed of the phishing campaign beforehand nor received any form of training for the detection of phishing email.

The campaign was launched on June 3rd 2019 at 11:00 AM. Phishing resources (e.g., the website) remained live for five working days. At `IND`, the phishing campaign was executed as part of a periodic security awareness training. The phishing emails were sent in batches, one for each experiment condition, every eleven minutes to avoid triggering server-side spam alerts. During the experiment we collect: `userId` (anonymized); `event` (email opened, website loaded, credentials submitted); `time of event` (timestamp). The `user category` (`Junior`, `Senior` or `Staff`) and `experiment condition` (treatment group) are reconstructed through the `userId`.

5.4.3. Experimental limitations

In our experiment, we added treatments incrementally to 'baseline' (treated) emails (e.g., $R^D \rightarrow R^D_{Au} \rightarrow R^A_{Au}$). Due to this choice, we cannot isolate the effect of 'multiple attacks' on a specific cognitive vulnerability, from the sole added effect of modifying the notification method. However, the alternative of applying treatments directly to R^D is undesirable because it will introduce *two* changes to the default email simultaneously (the cognitive attack *and* the notification method), making it difficult to distinguish the single contribution of each adopted persuasion technique. Similarly, as the implementation of a cognitive attack does depend on its position in text, simply applying the notification method to the baseline implementation would produce undesirable syntactic anomalies in the email bodies. In practice, this limits the set of meaningful comparisons we can make to R^D vs. the 'baseline' cognitive attacks, and to each 'baseline' cognitive attack vs. the one treated with the corresponding notification method. Finally, we consider a phish to be successful when a user submits their credentials. In reality, just visiting a webpage can compromise a user's system.

To minimize learning effects that may skew the results we limit our campaign to only one email per victim. This reduces the number of overall observations, but allows us to derive interpretable conclusions while not stretching ethical concerns on the experiment [293]. Whereas the final subject sample is relatively large, the adoption of nine experiment conditions limits the number of subjects assigned to each treatment to 83 overall. To evaluate statistical significance, we report robust confidence intervals for all results. However, as only a fraction of users fall for the attack, statistical tests across all user categories and experiment conditions may be deceiving. For this reason, we report descriptive statistics for all user categories and experiment conditions (Table 5.7), and run statistical tests only on aggregated results (Tables 5.8 to 5.10).

5.5. Results

Table 5.7 reports an overview of the submission rates by employee category and experiment condition. Overall, the observed success rate at `IND` is 29.9%, whereas the success rate at `UNI` is 15.4%; the overall success rate for both campaigns is 22.2%. The experiment conditions with the highest success rate belong to `IND`, with the exception of R^A_{Au} that scores higher for the `Junior` category at `UNI`. This suggests that, in general, `IND` is more vulner-

Table 5.7: Overall submission rates per experiment condition and category

Numbers **underlined and in bold** are the highest figures per experiment condition (rows). Numbers in **bold but not underlined** are the highest values within that organization. The overall success rate of the campaign is 22.2%; IND appears to be the most vulnerable with an overall attack success rate of approximately 30%. By contrast, the recorded success rate against the UNI group is 15.4%. The Junior category is in general the most vulnerable, but appears to be particularly susceptible to Authority attacks in the UNI group.

	Overall					UNI					IND				
		#Submissions (%)					#Submissions (%)					#Submissions (%)			
ID	no.	Overall	Junior	Senior	Support	no.	Overall	Junior	Senior	Support	no.	Overall	Junior	Senior	Support
R^{D}	83	19 (22.9)	11 (**27.5**)	5 (16.7)	3 (23.1)	44	5 (11.4)	4 (**17.4**)	0 (0.0)	1 (14.3)	39	14 (35.9)	7 (**41.2**)	5 (31.3)	2 (33.3)
R^{D}_{Au}	83	21 (25.3)	10 (25.0)	8 (**26.7**)	3 (23.1)	44	5 (11.4)	1 (4.4)	3 (**21.4**)	1 (14.3)	39	16 (41.0)	9 (**52.9**)	5 (31.3)	2 (33.3)
R^{A}_{Au}	83	19 (22.9)	11 (**27.5**)	6 (20.0)	2 (15.4)	44	11 (25.0)	7 (**30.4**)	3 (21.4)	1 (14.3)	39	8 (20.5)	4 (**23.5**)	3 (18.8)	1 (16.8)
R^{D}_{Lk}	83	20 (24.1)	9 (22.5)	8 (**26.7**)	3 (23.1)	44	6 (13.6)	5 (**21.7**)	1 (7.1)	0 (0.0)	39	14 (35.9)	4 (23.5)	7 (**43.8**)	3 (50.0)
R^{A}_{Lk}	83	19 (22.9)	12 (**30.0**)	5 (16.7)	2 (15.4)	44	8 (18.2)	6 (**26.1**)	2 (14.3)	0 (0.0)	39	11 (28.2)	6 (**35.3**)	3 (18.8)	2 (33.3)
R^{D}_{Sc}	83	16 (19.3)	8 (20.0)	7 (**23.3**)	1 (7.7)	44	4 (9.1)	3 (**13.0**)	1 (7.1)	0 (0.0)	39	12 (30.8)	5 (29.4)	6 (**37.5**)	1 (16.8)
R^{A}_{Sc}	83	23 (27.7)	14 (**35.0**)	5 (16.7)	4 (30.8)	44	9 (20.5)	7 (**30.4**)	0 (0.0)	2 (28.6)	39	14 (35.9)	7 (**41.2**)	5 (31.3)	2 (33.3)
R^{D}_{Cn}	83	15 (18.1)	7 (17.5)	5 (16.7)	3 (**23.1**)	44	7 (15.9)	5 (**21.7**)	1 (7.1)	1 (14.3)	39	8 (20.5)	2 (11.8)	4 (25.0)	2 (**33.3**)
R^{A}_{Cn}	83	14 (16.9)	7 (**17.5**)	5 (16.7)	2 (15.4)	44	6 (13.6)	4 (**17.4**)	1 (7.1)	1 (14.3)	39	8 (20.5)	3 (17.7)	4 (**25.0**)	1 (16.7)
Tot	747	166 (22.2)	89 (**24.7**)	54 (20.0)	23 (19.7)	396	61 (15.4)	42 (**20.3**)	12 (9.5)	7 (11.1)	351	105 (29.9)	47 (**30.7**)	42 (29.2)	16 (29.6)

Table 5.8: Odd ratios between [Senior, Support] and Junior categories

	Overall			UNI			IND		
Category	# (%)	OR	95% CI	# (%)	OR	95% CI	# (%)	OR	95% CI
Junior	89 (24.7)	–	–	42 (20.3)	–	–	47 (30.7)	–	–
Senior	**54 (20.0)**	**0.76**	**[0.54, 1.07]**	**12 (9.5)**	**0.41**	**[0.27, 0.63]**	42 (29.7)	0.93	[0.60, 1.45]
Support	23 (19.7)	0.75	[0.46, 1.22]	**7 (11.1)**	**0.49**	**[0.25, 0.68]**	16 (29.6)	0.95	[0.53, 1.67]

able to the attack On the other hand, attacks exploiting Authority seem more effective on Junior employees at UNI, supporting previous observations on the domain-dependent effect of cognitive attacks [347, 379].

The most effective cognitive attacks appear to be Authority and Scarcity (see 'Overall' column in Table 5.7), although we observe wide fluctuations between UNI and IND for Authority. For instance, for the Junior category R^{A}_{Au} appears to provide a large boost compared to R^{D}_{Au} in UNI (30.4% vs. 4.4%), whereas it *halves* the success rate in IND (23.5% vs. 52.9%). We observe a decreased success rate across all employee categories in IND for R^{A}_{Au} when compared to R^{D}_{Au}. By contrast, only the Junior group in UNI was impacted by the treatment, whereas more senior employees were not. The use of *subject line* for Scarcity (R^{A}_{Sc}) achieves higher success rates across both organizations, although by a lesser extent than R^{D}_{Au}.

RQ1: Does professional role have an impact on success rate of tailored phishing attacks?

The overview in Table 5.7 suggests that, in both organizations, the Junior category is the most vulnerable. Differences across other categories appear to be smaller. To more formally evaluate this, we report in Table 5.8 the relative change in submission rates related to the Senior and Support categories when compared to the Junior category. The overall figures show an only marginally significant negative effect for the Senior category (odds

Table 5.9: Odd ratios of Cognitive Vulnerability vs. Default treatment

	Overall			UNI			IND		
ID	**# (%)**	**OR**	**95% CI**	**# (%)**	**OR**	**95% CI**	**# (%)**	**OR**	**95% CI**
R^D	19 (22.9)	–	–	5 (11.4)	–	–	14 (35.9)	–	–
R^D_{Au}	21 (25.3)	1.14	[0.56, 2.33]	5 (11.4)	1.00	[0.27, 3.73]	16 (41.0)	1.24	[0.49, 3.09]
R^D_{Lk}	20 (24.1)	1.07	[0.52, 2.19]	6 (13.6)	1.23	[0.35, 4.38]	14 (35.9)	1.00	[0.39, 2.52]
R^D_{Sc}	16 (19.3)	0.80	[0.38, 1.69]	4 (9.1)	0.78	[0.20, 3.12]	12 (30.8)	0.79	[0.30, 2.04]
R^D_{Cn}	15 (18.1)	0.74	[0.35, 1.59]	7 (15.9)	1.48	[0.43, 5.06]	**8 (20.5)**	**0.46**	**[0.17, 1.27]**

Table 5.10: Odds ratios of Notification Method vs. Cognitive Vulnerability treatments

	Overall			UNI			IND		
ID	**# (%)**	**OR**	**95% CI**	**# (%)**	**OR**	**95% CI**	**# (%)**	**OR**	**95% CI**
R^A_{Au}	19 (22.9)	0.88	[0.43, 1.79]	**11 (25.0)**	**2.60**	**[0.82, 8.25]**	**8 (20.5)**	**0.37**	**[0.14, 1.01]**
R^A_{Lk}	19 (22.9)	0.94	[0.46, 1.92]	8 (18.2)	1.41	[0.45, 4.46]	11 (28.2)	0.70	[0.27, 1.83]
R^A_{Sc}	23 (27.7)	1.61	[0.78, 3.32]	**9 (20.5)**	**2.57**	**[0.73, 9.09]**	14 (35.9)	1.26	[0.49, 3.24]
R^A_{Cn}	14 (16.9)	0.92	[0.41, 2.05]	6 (13.6)	0.84	[0.26, 2.72]	8 (20.5)	1.00	[0.33, 3.00]

ratio of 0.76). Overall, belonging to `Senior` reduces the chances to submit the credentials by approximately 24% when compared to `Junior`. `Support` performs similarly, but the smaller sample size reduces its statistical significance.

Looking at results by single organizations, we find that `Senior` and `Support` in `UNI` are significantly less susceptible to the campaign w.r.t. `Junior` ($OR = 0.41$ and 0.49 respectively): `Senior` and `Support` employees at `UNI` are approximately 60% less likely to submit their credentials compared to employees in the `Junior` category. By contrast, the campaign success does not seem to vary significantly across professional roles in `IND`. On the other hand, we find that the `Senior` and `Support` groups are significantly less vulnerable at `UNI` than at `IND`, when compared to the respective `Junior`. This suggests that professional seniority and expertise may affect victims of phishing attacks differently across organizations.

RQ2: Are cognitive vulnerabilities effective in tailored phishing attacks?

The effect of cognitive vulnerabilities vs. the default email R^D is reported in Table 5.9. In contrast with previous literature [63, 155, 371, 379], we find no significant effect for the *mere introduction* of cognitive attacks in our phishing campaigns. The implementation of a *tailored* attack appears to reduce the effect of 'standard' cognitive attacks. The only marginally significant effect we observe *within* an organization is the negative impact of R^D_{Cn} in `IND` when compared to the effect of R^D. The addition of `Consistency` *decreases* the success rate of the default attack in `IND`. This result is also confirmed *across* organizations, whereby `UNI` shows a significantly higher vulnerability to `Consistency` than `IND` when compared to the effect of R^D.

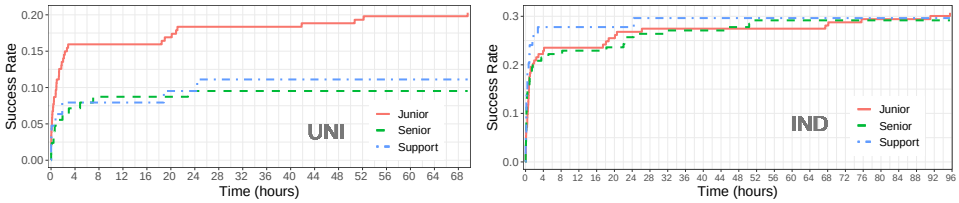


Figure 5.1: Success rate over time per user category in UNI (top) and IND (bottom). The grid lines mark 2 hour intervals.

RQ3: Does the use of notification methods to deliver cognitive vulnerabilities increase the success rate of a tailored phishing attack?

The effect of notification methods for the delivery of cognitive attacks is reported in Table 5.10. It is worth noting that, except for Scarcity (R^A_{Sc}), the effect of notification methods reverses across UNI and IND. For instance, the use of *contact information* for Authority (R^A_{Au}) has a positive effect in UNI ($OR = 2.60$) whereas it has a negative effect in IND ($OR = 0.37$), with the latter being marginally significant. Nonetheless, the difference across UNI and IND is strongly significant, and indicates a relative change in chances of success of $2.60/0.37 = 7.03$ times (i.e., approx. 603%), highlighting the highly-dependent effectiveness of the adopted treatment across the two environments. In contrast, the uses of *subject line* for Scarcity (R^A_{Sc}) has a positive effect on both UNI ($OR = 2.57$) and IND ($OR = 1.26$). However, the effect is not significant for IND and only marginally significant for UNI.

RQ4: Which characteristics of the campaign and victims affect the velocity of a tailored phishing attack?

Figure 5.1 shows the progress of the success rates over time for both organizations and by user category. We observe that victims' responses to the phishing campaign is almost immediate: for all categories in both UNI and IND, 50% of submissions occurred within the first 2 hours of the first day and approx. 75% of submissions occurred within 4 hours. We also observe that the victimization rate slows down as time passes across all categories, suggesting that most users that will fall for the attack will do so almost immediately.

5.6. Discussion

The effect of persuasion techniques in tailored phishing campaigns

The main outcome of our investigations is that, in tailored phishing campaigns, the means by which cognitive attacks are delivered to the victims appear to be more important than the mere presence of a cognitive attack itself. This suggests that, in presence of a tailored attack, 'cognitive effects' may be superseded by the overall 'persuasive' effect of a well-engineered phishing email. On the other hand, the means by which the cognitive vulnerability is delivered to the victim (i.e., its notification method) do have a sizable effect, although this does

not appear to be stable in direction across experiment conditions. This is particularly evident in the case of `Authority`, for which forged 'contact information' more than doubles the chances of success at `UNI`, but at the same time halves the chances at `IND`. This suggests that the notification method can be very effective in enhancing a cognitive attack, but it must be precisely tailored against the target – otherwise, it can backfire.

As notification methods move the cognitive attack from the body text of the phishing email to a more prominent position, they may either increase the persuasive power of the attack, or expose it to additional scrutiny from the user (the interested reader can refer to Section 3.3.1 for a more detailed discussion). For example, addressing a person by name (*Personalization*) in official correspondence may not be common practice in some organizations. In line with this observation, `Senior` subjects at `IND` were less likely to fall for the attack when administered R^A_{Lk} (19%), as opposed to the 'untreated' version of the same cognitive attack R^D_{Lk} (44%). Likewise, forging authoritative contact information to sign off a phishing email may backfire, particularly with `Senior` personnel: in `IND`, R^D_{Au} achieves a submission rate of 52.9%, which is essentially halved in R^A_{Au} (23.5%) once the forged contact information is added in the email signature.

Other effects appear to be more stable across organizations; for example, `Scarcity`'s effectiveness is magnified by modifying the subject line of the email both at `IND` and `UNI`. This suggests that modifying the delivery of cognitive exploitation methods may further improve the attack success rate. An alternative interpretation is that it could be the novelty of the implementation of the attack to make a difference in how easily can the attack be spotted.

Implications for research: The effect of notification methods in advanced, tailored phishing attacks is a little-explored area of research. Our results suggest that the effect of specific cognitive attacks (e.g., `Authority`) may be greatly enhanced by its positioning in a phishing email. This opens new paths forward in research on the relation between attack features, delivery and (expected) success. For example, this calls for new methods to evaluate *how* specific cognitive attacks are delivered in the wild (e.g., in multistage spear-phishing attacks [15]), and new experimental procedures to replicate these effects in laboratory settings. Similarly, this opens towards the development of risk-based detection tools looking for evidence of specific notification methods implementing cognitive attacks in phishing.

Implications for practice: Our results suggest that the implementation of consistent email designs through enforcement of internal policy may aid users in distinguishing 'legitimate' from 'illegitimate' requests. This is substantiated by the consistent trend whereby more experienced employees are more consistent in identifying 'anomalies' in the communication introduced by our treatments. Ideally, the devised internal policies could prescribe email structures and language that is *opposite* to that of an effective attack (e.g., no reminders in email subjects, name-only sign offs, etc.).

Susceptibility to tailored phishing

Our results illustrate that the professional role can have a significant impact on the success of a tailored phishing campaign. At the same time, they show that this effect can significantly vary across organizations operating in different domains. This is evident by comparing `UNI`'s

and IND's cases, where the Junior category appears to be significantly more susceptible to phishing than the Senior and Support categories in UNI, but no measurable effect emerges in IND. A possible explanation for UNI, in line with [52], is that Senior and Support can be considered as experienced users that are able to spot inconsistent patterns (or 'anomalies', as framed in Chapter 3), and potentially targeted phishing, due to their (on average) longer experience in the context of the organization itself. Surprisingly, this is not the case for IND, where the apparently more senior users score similarly to the less senior ones. A hypothesis is that the turnover rate is higher at IND compared to UNI for Senior and Support employees: newer employees in Senior and Support positions could lack experience in the organization processes and practices whilst being experienced professionals.

A surprising result is that the campaign against IND achieved much higher success rates than the campaign against UNI, despite IND employees periodically and structurally receiving phishing awareness training, whereas employees at UNI do not. This suggests that widely different results can be achieved across different domains and organizations. More studies are needed to systematically evaluate whether specific domains (e.g., education, manufacturing, finance) are vulnerable to specific types of attacks. This also indicates that previous phishing measurements relying on experiments with students and employees (see [320] for an overview) may be limited in validity across different application domains.

Implications for research: The evaluation of training effectiveness is an issue per-se [363]. The case of IND vs. UNI raises an interesting flag on the (lack of) interaction between training activities and susceptibility to phishing attacks. For example, it is known that training effectiveness decreases with time [31, 52], and may be less and less effective as users achieve higher awareness levels. However, whether the remaining fraction of users *cannot* be treated (i.e., they *will* fall for the attack, irrespective of the training received) is an open question in its own right. Further research on marginal returns of training activities could shed light on the overall effectiveness of awareness campaigns. Finally, our results call for replication studies evaluating the varying effect of phishing techniques against different user populations and organizational settings.

Implications for practice: Victims' lack of knowledge on internal organization processes can be exploited by an attacker to build credible pretexts and get a foothold within the organization (e.g., to perform lateral phishing attacks [148]). Junior staff may be most vulnerable to sufficiently tailored phishing attacks mimicking organizational settings. This may also extend to newly hired personnel in senior professional roles. Given this baseline of vulnerable users, specific training targeted towards these groups (e.g. based on URL detection [363]) may help reducing the attack surface.

Time characterization

Our time analysis shows that most users that will fall for the campaign will do so in its first few hours. The velocity at which phishing victims fall for the attack suggests that, when a tailored phishing campaign has been identified, chances are that a significant portion of targeted employees has already been compromised. This very short time makes phishing attacks dangerous and difficult to counter to date, since there is little to no time for defenders to react and contain the attack.

Implications for research: Whereas traditionally phishing prevention activities deal with *detection*, the findings above suggest that an effective *response* strategy may be more apt for tailored phishing. For example, the development of risk metrics for phishing attacks notified by highly-aware users in an organization could help mitigating the success of highly effective attacks within the first few hours of the campaign. This suggests that effective mechanisms for user reporting of phishing attacks may be a significant, yet untapped, resource to enable mitigation of advanced campaigns.

Implications for practice: Any effective reaction from an organization's response team should happen within the first few hours since the start of the campaign to be effective at all. Appropriate prioritization metrics, as well as response strategies, should be in place to address the threat in a timely manner [347]. Awareness from the response team of the ongoing attacks is key in this context. The competent departments at the organization should provide employees with clear instructions on how and when to report suspicious emails, have in place methods to promptly triage user reports and well-tested mitigation procedures ready to be deployed.

5.6.1. Threats to validity

Construct validity. As we cannot evaluate whether users read the email, treatment effect can only be assumed. Similarly, user categories and cognitive treatments might account for only a part of our outcome variables. Further, the pretext scenario may not reflect the real procedure for holiday hours at the organization. To mitigate this, we involved employees at the security department of both organizations since experiment design time.

Internal validity. Timing measurements are susceptible to unaccounted variables such as email load, working habits, and personality traits. Subject characteristics, like age, gender, nationality etc., may also have an impact on the results [256, 379]. The subject randomization at design time limits the statistical effect of these subject variables. Nonetheless, the sample size may affect the accuracy of the estimated effects. The designed treatments were designed to influence participants according to Cialdini's persuasion principles. There are inevitable variations in the users' perception of these manipulations. We mitigated this risk by implementing cognitive attacks based on the findings of previous studies (cf. Sec. 5.4.1). Finally, as we do not check for credential correctness, not all submissions may correspond to a successful phish.

External Validity. The sample encompasses academic staff and industry workers, which can have different characteristics than, e.g., private users or clients of a bank. Similarly, our results may not generalize to other organizations and types of phishing (e.g., clone-based phishing, for which cognitive attacks may be less important).

5.7. Conclusion

In this chapter, we evaluated the efficacy of a tailored phishing scenario and the effects of persuasion techniques, their delivery method and context of deployment on the attack success rate. The tailored phishing attack against employees yields a high success rate, between

10% and 30% across user roles and organizations, thus highlighting the relevance of context in explaining attack success variability (e.g., academia vs. industry). Our experiment shows that an advanced target tailoring of the campaign overcomes the effects of more conventional, cognitive attack techniques. In particular, our results reveal that the means (i.e., in our experiment, the notification method) by which the cognitive attack (i.e., the persuasion technique) is delivered matters, and can further enhance a well-tailored attack against an organization. Finally, the measurement of attack velocity reveals the key role of time in the impact of a phishing attack, providing insights for more effective attack response practices.

Our results provide an answer to RQIII and stress the importance of replicating previous studies on phishing across different domains to evaluate how effects may vary in different organizational settings. The results and recommendations reported in this chapter suggest that further research on advanced, tailored phishing attacks is needed to support organizations' defensive processes and capabilities. In the following chapters (Chapter 6 and 7), we draw from the above conclusions to continue our investigation on RQIII by exploring technological and organizational mitigation strategies against phishing attacks.

6

Detecting zero-hour phishing web pages: visual similarity and user intervention

From the results of the previous chapter, we find that time is of the essence in reacting to (tailored) phishing attacks: the delivered URLs link web pages that victimize the targeted user base in the first few hours since delivery. Typical detection approaches, such as blocklists, are limited against these *zero-hour* phishing attacks due to the unknown features of new attack instances, such as the URL or page contents; detection evasion techniques employed by attackers, such as embedding targeted brand references in pictures, further complicate the identification of the phishing attempt. Some visual similarity-based detection approaches can overcome evasion techniques and spot an unknown phishing page by comparing the rendered content of a page with a known reference brand. However, such methods are limited by the need of a predefined reference list of domains or brand representations, such as logos and screenshots, hindering the defense possibilities of less resourceful organizations. Following on RQIII, this chapter demonstrates a technological mitigation strategy against zero-hour phishing that relies on search engines to identify which website a phishing page is replicating by means of textual and visual features extracted from an unknown page. The tool, implemented as a browser extension, supports users in deciding on a website legitimacy with various risk communication methods. We evaluate the tool effectiveness in detecting phishing pages using a dataset sourced from phishing aggregators and suggest new research directions to investigate the complex interaction between users and decision support tools.

This chapter is originally published as B. van Dooremaal, P. Burda, L. Allodi, and N. Zannone, "Combining Text and Visual Features to Improve the Identification of Cloned Webpages for Early Phishing Detection", Tn the *16th International Conference on Availability, Reliability and Security* (ARES), ACM, 2021, pp. 1–10
and as P. Burda, L. Allodi, and N. Zannone, "A Decision-Support Tool for Experimentation on Zero-Hour Phishing Detection", In *Foundations and Practice of Security* (FPS), vol. 13877, Springer LNCS, 2022, pp. 443–452

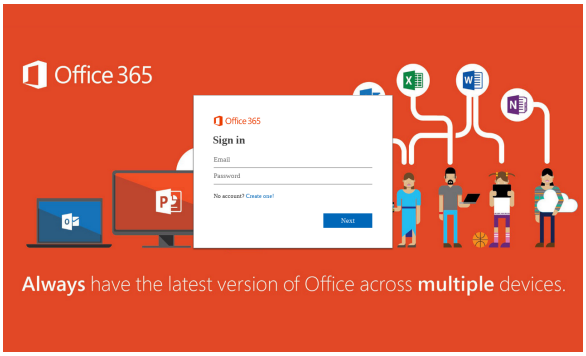


Figure 6.1: Phishing website imitating Microsoft Office 365 with no textual reference to Microsoft or Office in the web page DOM.

6.1. Introduction

Throughout the previous chapters we witnessed the emergence of increasingly sophisticated phishing attacks [15, 127, 232]. The adoption of innovative detection evasion techniques and the velocity at which phishing attacks arrive and change form make it challenging to design early detection systems able to warn users of the suspicious nature of a visited website. As the tailored phishing campaign of Chapter 5, new and unknown attack instances take their toll in the first few hours since delivery (hence *zero-hour* phishing), and attempt to bypass detection systems by fingerprinting user agents or concealing features of cloned pages in embedded objects, while preserving the visual similarity needed to persuade the end user they are indeed on the legitimate webpage [255].

Considerable research has been conducted in the domain of phishing detection, for example, block-listing phishing domains [128], bag-of-words filtering, visual similarity comparisons [163] and machine learning techniques [3, 213]. The timeliness and accuracy of these approaches, however, is limited; for example, updating a block-list of phishing domains is often too slow to match the velocity of an unknown attack and, filtering is often subject to a high misclassification rate. Visual similarity-based methods rely on accurately finding a corresponding legitimate page and can be limited by certain evasion techniques and, importantly, by the need of a predetermined list of brands to protect.

Among existing phishing detection approaches, some visual similarity-based approaches have the ability to support the detection of 'zero-hour' phishing websites [136, 163]. To identify the web page that the phishing page is most likely mimicking, these approaches typically extract features from the web page Document Object Model (DOM) and use these as search terms in search engines. Still, these efforts are limited by detection evasion techniques such as the replacement of text and logos by other objects (such as images and other embedded objects). This is, for instance, the case of a recent phishing website imitating Microsoft Office 365's login page, shown in Fig. 6.1, where any appearance of the brand names is in embedded objects, thus making the use of current feature extraction techniques ineffective.

In this work, we propose a method aimed at improving the automated identification of a

cloned web page by extracting visual features from the screenshot of a suspicious web page, in addition to textual features extracted from the DOM of the web page. Our approach exploits color contrast within portions of a screenshot to recognize regions of interest on the web page, which are then used as search terms for a reverse image search, thus following the common assumption of visual similarity-based approaches that search engines place benign results at the top [211]. The top results of the combined text search and reverse image search are deemed as the potential targets and compared with the suspicious page using image similarity metrics for phishing detection.

As seen in Chapter 2, recent studies suggest user awareness alone remain ineffective against the increasing sophistication of phishing attacks [28, 88, 272, 307]. Hence, the solution space for effective measures for early-detection of phishing websites is moving towards mixed approaches where technology and automation can support the decision-making of a (threat-aware) human. Yet, users are often not considered in the design of the tools themselves [6, 213]. Humans may not heed the generated warnings due to, e.g., lack of trust in the decision support system or additional user interface fatigue, with consequent detrimental habitual patterns [235]. Moreover, the amount, type and even content of warnings can depend on the employed detection system and, consequently, affect warning effectiveness. In this scenario, extant phishing detection tools are often limited in their applicability to experiments that capture the full process where the interaction between the phishing web page and the user unfolds. For example, even the best detection methods can be ineffective when users do not trust and follow the tool's advice [70]. On the contrary, pitfalls of detection tools, such as false positives or long run-times, can be mitigated with effective risk communication. We argue that these limitations narrow the research possibilities where technology and automation can support individuals in avoiding phishing attacks. Therefore, our approach yields an integrated phishing detection solution in the form of a client-server architecture where the detection logic stays on the server while the client is implemented as a browser extension that warns users when they visit a potential phishing page. Our proposition results in an *integrated research approach* that puts both phishing detection and Human Computer Interaction (HCI) ingredients together for an experimental tool to evaluate, characterize, and refine the interaction between phishing decision support, and the final user.

Our contributions towards answering RQIII can be summarized as follows:

- We devise a novel approach to identify which website a phishing web page is imitating by means of *both* textual features extracted from the DOM and metadata of the page and visual features (regions) extracted from a screenshot of the page.
- We evaluate our approach for target identification against a corpus of phishing attacks gathered from largely used anti-phishing sites such as OpenPhish, PhishTank and PhishStats. The results show that our approach achieves an accuracy of 99.2%.
- We provide insights on the effect of target identification through visual features on the classification of phishing web pages. Our experiments show that the use of visual features, in addition to textual features, reduces the misclassification rate by 67%.
- We discuss how our tool can be used to instrument scientific research towards an integrated research line on zero-hour phishing allowing, for instance, the evaluation of user trust in detection tool's advice or the exploration of new risk communication methods by

keeping track of past decisions and associated risks.

This chapter is structured as follows: Section 6.2 provides background on automated phishing detection and target identification techniques. Section 6.3 presents the overall methodology with a particular focus on target identification. An evaluation of the proposed methodology is presented in Section 6.4. Section 6.5 discusses our results and ways forward for research. Finally, Section 6.6 concludes the chapter.

6.2. Background and Related Work

6.2.1. Website phishing detection

Previous work on automated detection of phishing websites can be grouped into three main classes, namely list-based, visual similarity-based, and heuristic-based approaches [163].

List-based approaches operate by comparing the universal resource locator (URL) a user visits against a list of known phishing websites (a so-called *block-list*) or a list of legitimate websites (a so-called *allow-list*). These lists are typically maintained and updated by relying on external sources such as PhishTank [279] or Google Safe Browsing [128]. List-based approaches are, however, ineffective, especially against zero-hour attacks. Previous work [272] shows that several phishing websites are hosted on compromised domains, which can affect list-based approaches (especially based on allow-lists). Additionally, solutions based on a block-list can only detect a phishing website if that website is present in the list. Therefore, phishing websites have to be detected through other means before they can be detected with this method. Previous measurements [143, 255] observed that Google Safe Browsing and PhishTank can take from 9 hours to twenty days to add a URL to their lists since a phishing attack has been seen for the first time, making list-based approaches largely ineffective against zero-hour attacks. On the other hand, heuristic-based and visual similarity-based approaches can detect zero-hour phishing attacks [136, 163].

Heuristic-based approaches analyze features extracted from a web page using predefined rules to determine if the web page is legitimate [163]. Due to advancements in phishing attacks many features that heuristic-based approaches rely on are unreliable for classifying a website. For example, the presence of SSL certificates, which was historically used to classify the legitimacy of web pages [263], is no more a valid heuristic [28]. In general, attackers can forge relevant features to be invisible to the heuristic rules, but still visible to the user [163].

On the other hand, visual similarity-based approaches use content rendered in the web browser to determine the (non)legitimacy of a website. These techniques use features such as the favicon (a small image next to the website title) [72], the logo on the web page [6], or a screenshot of the entire web page to compare two websites and determine whether one is attempting to imitate the other [6, 210, 213]. The advantage of visual similarity-based techniques over heuristic-based approaches is that the replacement of text by other objects (such as images, Flash, and other embedded objects) by the attacker cannot circumvent the detection technique [315]. However, the ability of visual similarity-based approaches to detect new instances of phishing attacks depends on their ability to find the legitimate website a phishing web page is imitating.

Table 6.1: Studies that attempt to identify the target of a phishing web page. Accuracy is self-reported by the authors.

Authors	Accuracy	Feature type	Language Dependency	Description
Peng et al. [272]	88.0-99.8%	Text + OCR	Dependent	Uses OCR on a screenshot of the web page. Filters text found with RAKE and searches based on that.
Ding et al. [92]	99.6%	Text	Independent	Uses the title of the web page as search terms.
Ramesh et al. [288]	99.6%	Text	Dependent	Uses term frequency-inverted document frequency on the corpus of the English language and text extracted from title, meta tags, and the body of a web page to form search terms.
Muppavarapu et al. [243]	98.5%	Text	Independent	Extracts title and keywords from web page and uses information distance algorithms to form search terms.
Marchal et al. [219]	97.3%	Text	Independent	Extracts text from multiple fields and uses iterative queries to a search engine to refine search terms
Wenyin et al. [370]	92.1%	Links	Independent	Uses crawler to build a directed graph. Uses search engine based on title of all pages in the graph
Liu et al. [211]	91.4%	Links	Independent	Generates a cluster based on the web page and associated pages. Compares cluster to known clusters
Chiew et al. [72]	95.5%	Visual	Independent	Reverse image searches using the Favicon of a website.
Chiew et al. [71]	87.0%	Visual	Independent	Uses machine learning to identify which image resource is the logo and reverse image search that image.
Adebowale et al. [5]	98.3%	Hybrid	Independent	Uses machine learning with text, frame and image features and various legitimate sources, like internet articles and trusted lists.

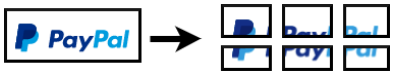


Figure 6.2: Example of splitting an image of the PayPal logo to evade detection when using resources directly.

6.2.2. Targeted brand identification

A few visual similarity-based approaches identify the candidate pages for comparison from a predetermined corpus of benign web pages [6, 210, 213, 229, 387]. This, however, hinders the detection of zero-hour phishing attacks or any attack imitating a domain not present in the given corpus, similarly to list-based approaches. Therefore, the identification of the page resembling the page under analysis can be performed using search engines. Specifically, visual similarity-based approaches often apply keyword extraction methods to extract relevant terms from the web page, which are then fed to a search engine. The assumption underling these approaches is that the website of targeted brand is more popular than the phishing website imitating it and that search engines place benign websites on top. Table 6.1 presents an overview of the techniques used to extract search terms from a web page[1].

As shown in the table, many text-based algorithms extract search terms from the title-tag and other metadata of the web page [92, 243, 288]. Liu et al. [211] propose to identify the targeted brand based on the hyperlinks on the phishing website and keywords from the web page. However, the brand name cannot be detected using Liu's or any text-based approach when the brand name occurs only in embedded objects and images, as it is the case for the web page in Fig. 6.1. To alleviate these limitations, Peng et al. [272] propose to use a screenshot of the web page, instead of the hypertext markup language (HTML) content, in combination with optical character recognition (OCR) techniques to extract the visible text. The text mining algorithm RAKE (Rapid Automatic Keyword Extraction) [300] was used to remove less important words, and the remaining terms were searched on Google. The authors observed a notable accuracy variance between popular targets such as PayPal and Microsoft where an accuracy of 99.8% was achieved and less known brands where the accuracy dropped to 88.0%. Adebowale et al. [5] propose to use features extracted from both the text and images. However, the proposed approach does not use the visual properties of a web page, but only the DOM elements related to images in the page [5], thus failing to extract the brand names from the phishing website of Fig. 6.1.

Reverse image search has been employed to retrieve the origin of image resources (favicon [72] or images deemed to be the logo by a machine learning algorithm [71]) when textual features are insufficient. However, extracting the images used in the website might be still not sufficient to identify the targeted brand. For instance, images can be decomposed into several parts and then displayed as if they belong to a single image; a reverse image search on each part might not allow finding the legitimate page. Fig. 6.2 shows an example of splitting image resources. Moreover, an adversary could convert images to pure cascading style sheets (CSS) [64] which results in no resources for reverse image lookup.

To understand the impact of the increasing sophistication of phishing attacks on existing

[1] *Language independence* in Table 6.1 refers to the property that the approach does not rely on or use properties of a language to find its origin, e.g. no bag-of-words or stop-word filtering.

techniques, we performed a replication study of the approach proposed in [92][2] using a dataset comprising 135 phishing websites and 396 benign websites. The phishing samples were collected from aggregators such as PhishTank and OpenPhish in September 2019. We observed that the accuracy dropped to 86.3% for phishing websites and 90.3% for benign websites compared to the 99.6% accuracy reported by the authors. The main cause for this discrepancy is the lack of references to the brand name in the title or text of phishing web pages, which often contains only simple sentences such as "Login To Continue". Additionally, we observed that performing text mining on the DOM of phishing web pages often does not provide information on the targeted brand as in the case of the web page in Fig. 6.1.

Although previous studies [163] have noted that the target identification accuracy is highly reliant on the underlying feature extraction algorithm and search engine, the concept of feeding features to a search engine remains a powerful tool to identify the website a phishing website is imitating and eliminates the need to maintain an updated list of benign websites. Nonetheless, a more robust way of extracting search terms, which goes beyond applying text mining on HTML tags or images extracted from the web page for a reverse image search, is required to prevent evasion techniques as the ones of Figs. 6.1 and 6.2. In the next section, we present an approach to target brand identification that uses both textual features and visual features extracted from the screenshot of a phishing web pages, making it the first solution to effectively combine both.

6.2.3. Warnings

Warnings are the primary means to communicate security risks to users [235]. Two main categories are often employed in web browsers: *passive warnings*, which warn users without blocking the content area of a webpage, and *active interstitial warnings*, which block the content area and require an active interaction from the user to be bypassed [181]. Active warnings are more likely to be heeded by users than passive ones and, therefore, considered more effective in averting phishing attacks. Nonetheless, more experiments are needed to understand the effects of these warnings, which still suffer clickthrough rates between 9-18% for the phishing warnings and up to 70% for SSL related warnings [10]. Active warnings carry the risk of disrupting applications' usability too often, to a point where users can develop habitual and detrimental behavior patterns (such as overriding security settings), nullifying warning effectiveness altogether [235]. Moreover, user compliance is very sensitive to the context where warnings are triggered; for example, higher compliance was observed in online banking than in an e-commerce context [70]. Recent work has investigated how to nudge users to pay attention to warnings, for example, with *just-in-time*, *just-in-place* tooltips that elicit a more systematic cognitive response without blocking users completely [362]. This recent line of research integrates multiple disciplines and yields promising results, further signalling the need of new and innovative experimentation in this direction. Overall, research on warnings tends to disregard the internal mechanisms of phishing detection methods. On the other side, users are often not considered in the design of such methods. This has the side effect of limiting methods and tools' applicability to experiments that capture the full process where the interaction between phishing webpages and users unfolds.

[2]We selected the approach in [92] due to the highest reported accuracy among the works in Table 6.1.

6.3. Approach

To enable experimentation in the context of early phishing detection, we designed and developed a tool that employs a visual similarity-based phishing website detection method as the backend and leverages a variety of warning mechanisms to inform the user about the identified risks posed by the webpage they are visiting. An overview of the overall tool architecture is presented in Fig. 6.3. The tool is available at: `https://github.com/paolokoelio/zerohour-decisionsupport-phishing`.

6.3.1. Overview of the approach

The backend of our tool implements the machinery for early detection of phishing websites, which operates on a remote server exposing a REST-like API. The duty of the backend is to evaluate the legitimacy of a given web page (e.g., from the perspective a user's browser) by relying on textual and visual features of the web page. Fig. 6.3 (Backend) illustrates how the approach operates. By relying on search engines and the visual features of a rendered webpage (rather than only on features of the DOM), the approach allows a zero-hour protection by avoiding the maintenance of benign allow-lists, and is robust against resource evasion techniques, such as image splitting (Fig. 6.2) or the replacement of image resources by pure CSS. In addition, the approach uses information that the attacker deploying the phishing page gives in input to their targets rather than other hidden factors, such as the source code. Therefore, if an attacker attempts to distort an image to evade detection, it is equally visible to its target (who would then distrust the web page, foiling the attack [3]) as it is to our approach.

The backend takes as input a website and obtains the DOM and a screenshot of the rendered web page, after a run of all scripts (1). The retrieved web page and its screenshot are used to extract textual and visual features for target identification (2). Textual features are extracted from the DOM (e.g., HTML tags) in a similar fashion to the techniques in Table 6.1. In addition, regions potentially containing identifiable information are extracted from the screenshot. These features include, but are not limited to, logos, slogans, parts of header images, and other visual information that is likely to be found in the corresponding legitimate website (cf. Sec. 6.3.2).

The extracted features are used as search terms to find websites similar to the current web page through a search engine (3). Hereafter, we refer to these web pages as *associated pages*. The intuition is that search engines most likely place benign results at the top [211]. Therefore, if the visited URL is within the top results, it is deemed to be legitimate. Relying on a search engine to find the associated pages guarantees the language independence of the approach and zero-hour protection as well as it avoids the need of maintaining an updated list of all possible benign websites.

Textual features are used as search terms in a text-based search engine (e.g., Google, Yahoo, Yandex), whereas the portions of the screenshot identified as regions of interest are fed to a reverse image search engine. The top results of both searches are marked as candidate associate pages and used to determine whether the current page is legitimate or not. In our experiment (Sec. 6.4), we consider the ten top results of each search as potential associate pages.

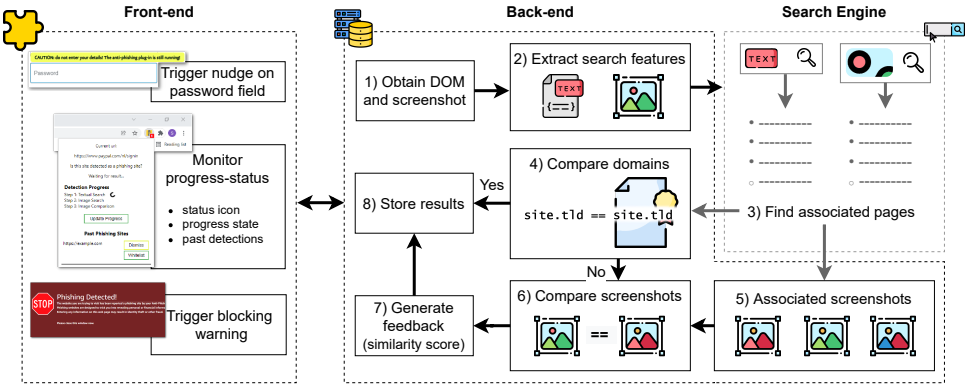


Figure 6.3: Components of the anti-phishing approach

To determine the legitimacy of the current website, its URL is checked against the domain names listed in the 'Subject Alt Names'-field of the associated pages' SSL certificates (4). This field contains all domain names for which the certificate is applicable. This is particularly useful for websites that have multiple domains/subdomains or languages, e.g. Amazon has multiple domains such as `amazon.co.uk` and `amazon.com`. Accordingly, while a search engine might accurately determine the brand, it might not return the exact domain or language version. If the URL is in this list, the website is marked to be legitimate.

If the domain name of the current web page does not appear in the SSL certificate of any candidate associated page, a screenshot of the associated pages is obtained (5). Each screenshot is visually compared with the screenshot of the current web page (6) and, depending on their degree of similarity, it is classified as 'phishing' or 'legitimate'. The similarity scores are then used to generate feedback (7) about the legitimacy of the current web page.

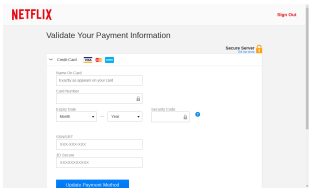
The feedback is stored in a database (8) and sent to the frontend (Fig. 6.3, Frontend) which modules are responsible for providing information on the current status of detection and triggering alerts.

6.3.2. Search feature extraction

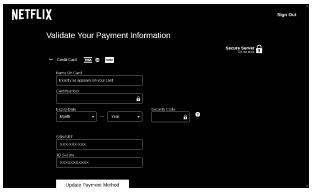
This section presents our approach to extract features from the DOM and the screenshots of a web page, which are used to find the corresponding associated pages.

Textual features

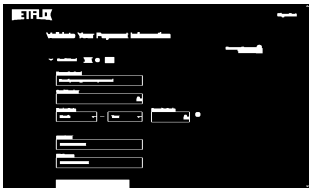
The text-based phishing detection methods previously discussed yielded a lower accuracy with the current phishing threat landscape than originally reported therein (cf. Sec. 6.2.2). However, the high speed and low-cost of text mining compared to image processing makes it an ideal preliminary source of target candidates. In addition, the use of textual features provides an ideal augmentation to the visual features for those websites containing little to no



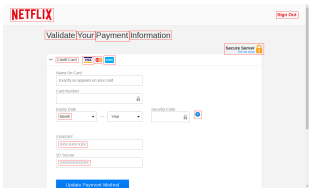
(a) Input image screenshot



(b) After using the thresholding algorithm (step 1)



(c) After performing morphological closing (step 2)



(d) After performing the analysis and filter (step 3 & 4)

Figure 6.4: Steps undertaking while extracting visual features from a screenshot of a phishing website impersonating Netflix.

6

image regions that allow inferring the origin of the website. In our experiments, we only consider the title of the HTML page; however, other textual features could be included as well.

Visual features

For the extraction of visual features, the underlying idea is, given the screenshot of a web page, to identify regions on the image that contain identifiable information about the legitimate website being impersonated. These regions will, when supplied to a reverse search engine, result in a list of pages that have visually similar regions. As a result of this procedure, it is important that the number of regions found is not too large due to the cost of querying a reverse image search engine. The steps undertaken for extracting relevant regions are as follows (Fig. 6.4 exemplifies):

Step 1: Websites typically place their logo in such a way that it is distinctive from the background. Following this intuition, we convert the screenshot of the web page to a black and white image to find areas with high contrast. To this end, we first obtain a grayscale version of the image and then apply Otsu's thresholding algorithm [262] on the obtained grayscale image. This algorithm takes a grayscale image as input and, based on the histogram of the image, transforms it into a black and white image. An example of the application of this step on Fig. 6.4a is shown in Fig. 6.4b.

Step 2: We perform morphological closing [310] on the black and white image such that nearby regions are connected. This is done by first manipulating the image such that the boundaries of white portions are expanded. This expansion connects nearby elements to each other, e.g. letters are connected to become a word. After the white portions are expanded, a shrinking technique is applied to remove the boundary expansion in directions

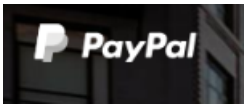


Figure 6.5: Light-colored PayPal logo on a darker background.

where no connection occurred. Fig. 6.4c illustrates the result of this step.

Step 3: Topological structural analysis [328] is then applied to the image to identify all connected components. Applying this algorithm gives us information about the boundaries and hierarchy of every region within the image.

Step 4: The previous step identifies all regions in the image with a notable contrast with the background. However, this includes regions representing text samples, buttons, and other items of the web page, which have a low likelihood of retracing back to the corresponding legitimate website. Therefore, we filter the identified regions based on heuristics to keep only those regions that potentially contain identifiable information (cf. Sec. 6.3.2).

After *step 4* a set of candidate regions that potentially contain identifiable information about the web page is obtained, as illustrated in Fig. 6.4d. It is worth noting that the morphological closing in *step 3* only enlarges and connects one type of region, the ones in white on Fig. 6.4c. This means that on a single pass the algorithm is able to identify darker logos on a light background. To recognize light colored logos on a dark background, e.g., Fig. 6.5, *steps 1-4* are reapplied by inverting the colors in *step 1*.

We can observe in Fig. 6.4d that the Netflix logo was correctly identified and extracted from the web page along with other regions of less interest. Tweaking on the strictness of the heuristics filter can tune the number of potential relevant regions identified. However, this can lead to excluding regions that are the sole identifier of the web page. At the cost of language independence, the regions resulting from *step 4* can be further reduced by performing optical character recognition in combination with a bag of words to exclude.

Region filtering

During visual feature extraction, a filter is applied (*step 4*) to remove regions that are unlikely to reveal the true identity of the page, i.e., reverse image search on these regions are likely not to return the page that is being imitated, which decreases reverse image searches on potentially irrelevant regions. We build such a filter based on a number of regions' properties: *height*, *width*, *area*, *number of colors*, *dominant color percentage*, and the *coordinates* on the web page.

The *height*, *width* and *area* are predominantly used to exclude regions that are likely to be text, e.g., regions with a height smaller than 20 pixels typically contain text that unlikely is identifiable for the web page. Additionally, we restrict the maximum size of regions because our experiments showed that large regions are often not representative of the website and yield low accuracy.

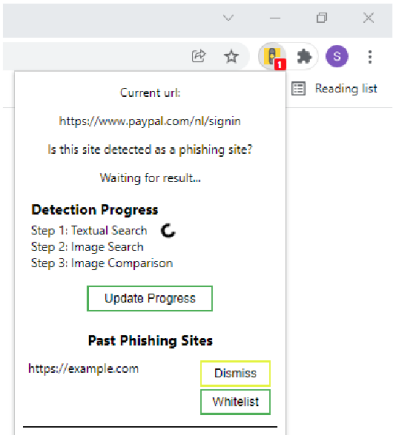


Figure 6.6: Extension status pop-up (past phishing sites are displayed)

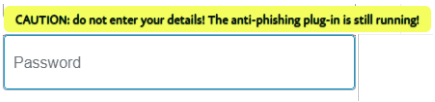


Figure 6.7: Passive, just-in-place tooltip when selecting a password field

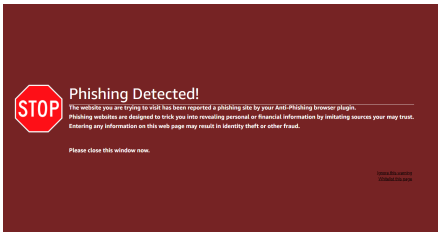


Figure 6.8: Active, full-screen blocking warning upon a successful detection

The *number of colors* and *dominant color percentage* can also be used to filter out regions that are likely to be text, which is typically characterized by the use of a few colors used and a high dominant color percentage [385]. It is worth noting that, although the Netflix logo in Fig. 6.4d looks monochromatic to the human eye, the number of unique colors (shaded variations of red) is actually in the hundreds, while the text regions beneath contain just tens of unique colors. On the other hand, an upper bound on the number of colors can be used to exclude regions representing portions of the background, like pictures, which are typically characterized by thousands of colors but do not provide identifiable properties.

The last set of region properties considered for filtering are the *coordinates* of the region. The sheer majority of phishing and legitimate websites in our datasets have meaningful regions concentrated within certain coordinate sets, like top-left or central parts of a web page. Regions of the screenshot that contain identifiable information hardly occur in other coordinate sets, e.g., the bottom right of the web page unlikely contains identifiable information. For web pages where many regions are found, priority can be assigned to coordinate sets that are more likely to contain identifiable information.

The filter evaluates each region against each property individually. Regions where at least one of the selected property values does not fall within a threshold are discarded (cf. Sec. 6.4.3 for an evaluation of the region properties used for filtering).

6.3.3. Frontend

To enable experimentation in which the human is in the loop, we realized a frontend interface as a Chromium browser extension. This allows to include HCI factors in the experimentation ingredients and facilitate experiment deployment. The modules of the plug-in are shown in Fig. 6.3 (Frontend).

The frontend extension is configured to only scan pages that contain a password field, given the focus on phishing websites aiming to steal user credentials. Upon visiting a page, the

detection process starts in the background (cf. Section 6.3.1), and a traffic light icon in the address bar signals the current status, as shown in Fig. 6.6. The user may click on the icon for more details. On the top, the current URL is displayed together with the outcome, the center contains information on the progress, i.e., the textual/image search and image comparison steps, and the bottom shows the past phishing discoveries.

Whenever users select the password field, the extension triggers the just-in-time just-in-place passive warning in Fig. 6.7 to remind that the detection in not complete. Researchers can personalize the warning behavior to steer user attention with different designs or impede certain actions by, e.g., temporarily blocking the "Submit" button. When a webpage is detected as a phishing webpage, a full-page blocking warning (Fig. 6.8) blocks the user if they are still on that page, akin to current browsers' behavior. To ignore this active warning or remember this choice the user must locate and click on the respective links. Contents and design of the message can be customized to, for example, embed information on the used search features or alter interaction paths to dismiss the message.

To cover cases where the user acts on the webpage before the analysis is complete, past phishing websites are displayed in a retrospective fashion, as shown at the bottom of Fig. 6.6: even if the user navigates away from the not-yet-detected phishing page, the system will alert the user retrospectively in the status icon and in the pop-up of a detection. Users can dismiss or label as "legitimate" a previously detected phishing URL. The displayed information and interactive elements of the pop-up can be altered to give less or more insights and control to the user, such as near real-time data, history of detection or (de)activation of features.

6.4. Evaluation

This section presents a performance evaluation of the target identification method presented in Section 6.3. Additionally, the impact of the improved target identification is evaluated on a phishing classification task by means of visual similarity classifiers.

6.4.1. Experimental Setup

We implemented a proof of concept of our approach for target identification in Python. Given a URL, the application uses a Docker container to visit the web page using a monitored Chrome browser and retrieves a screenshot and the DOM. The BeautifulSoup library [295] is used to parse the DOM and extract textual features. All image processing algorithms used in our approach (steps 2 and 6 of Fig. 6.3) are implemented with the OpenCV library [49].

To find the associated pages using visual features (step 3 of Fig. 6.3), we initially considered Google, Yandex and TinEye. However, Yandex and TinEye have a rate-limit of less than 200 searches per day, which made experimenting impractical. Hence, our application relies on Google as the search engine.

6.4.2. Data collection and sanitization

Public phishing datasets used in previous work (e.g. [2, 236, 330]) are restricted in size and are limited to features that are either too few, not robust (like reliance on SSL certificates),

Table 6.2: Target composition of the phishing dataset.

Brand	Entries	Percentage
PayPal	326	32.6%
Microsoft	116	11.6%
Apple	106	10.6%
Facebook	86	8.6%
Google	77	7.7%
Netflix	54	5.4%
Bank of America	46	4.6%
Cox	36	3.6%
Other	153	15.3%

or simply outdated. Also, common sources for benign websites, e.g. Alexa's most visited web pages [24], can suffer from bias [309]. Therefore, we built a new dataset of phishing and legitimate web pages to evaluate the performance of the proposed approach.[3]

Phishing dataset

The set of phishing web pages originate from 100,000 URLs that were posted in phishing feeds such as OpenPhish [258], PhishTank [279], and PhishStats [278] between September 2019 and December 2019. These are community efforts to aggregate URLs corresponding to phishing web pages. The collected URLs had a screenshot and their DOM stored. However, many of the web pages posted in these feeds have a short lifespan. To remove web pages that were taken offline, the DOM was checked against terms that are common in web pages that are no longer online, e.g. 'Error 404 - Not Found'. In this case the page was removed from the dataset. We also removed duplicate attacks from multiple time periods or domains by grouping pages with identical DOM and, thus, tested similar attacks only once. This is to avoid the dataset over-representing certain attacks. The remaining pages were randomly sampled to construct a phishing dataset of 1,000 entries. This dataset had its targets manually verified to ensure accurate labeling of the dataset. The composition of the dataset is shown in Table 6.2. The dataset contains multiple languages such as English, Portuguese, Russian, and Japanese. Most non-English phishing web pages attacked international brands such as PayPal's localized web site. A notable exception is Magazine Luiza, a Brazilian retail web site, which is exclusively in Portuguese and has 24 (2.4%) entries.

Benign dataset

The entries of the benign dataset originate from the DMOZ database [94]. This database contains 3,861,202 (assumed) legitimate web pages in 90 languages and is the result of a community project aiming to create an open-content directory of the Internet. The main advantage of using the DMOZ database over other sources for legitimate web pages often adopted in phishing research (e.g., the Alexa top 100 web pages [24]) is that this database contains also less popular web pages. Our approach relies on the performance of the employed search engine. Popular websites, such as the ones in the Alexa top 100, have dedicated search engine

[3]The dataset is available at `https://doi.org/10.5281/zenodo.4922598`.

optimization and search engines sometimes increase their ranking in the results. Therefore, it is important that less known websites are included as well, to ensure that the approach can be generalized to real world scenarios. However, while the DMOZ database only contains legitimate entries and is composed of a mix of popular and less known websites, it has not been updated since 2017. Therefore, to avoid possible redirects, we verified that the domain name of the landing URL matches the domain name of the URL being sampled. Additionally, as done for the phishing dataset, we check that the landing page is not an error page. The benign dataset comprises 1,000 randomly sampled domains that met these constraints.

6.4.3. Target identification performance

To assess the performance of target identification, we used the datasets presented in Section 6.4.2. Of the 2,000 websites (1,000 legitimate and 1,000 phishing), we used 70% for training and optimal filter parameters setup and the remaining 30% to validate the results.

Target matching

Our approach uses the results from two distinct sources, textual features and visual features, to find the associated pages. Table 6.3 presents the percentage of tested web pages that had the target of the web page correctly identified. For phishing web pages this means that the page being imitated was within the result set. On the other hand, for benign web pages, it represents the percentage of benign websites that are included in the certificates of one of the associated pages. The 'combined' row represents the accuracy when the associated pages obtained using both textual and visual features are considered. Table 6.3 shows that the accuracy of identifying the correct target using only visual features is 78.6% for phishing web pages and 72.3% for benign web pages. This is worse than using textual features, whose accuracy is 86.3% and 90.3% respectively. However, the use of visual features prove to be a good augmentation for the websites that avoid brand names in the DOM. Using both text and visual features improves the identification accuracy to 99.2% for phishing and 92.0% for benign web pages. This is comparable to the accuracy self-reported by other works (ref. Table 6.1) but on more recent phishing attacks. This will be discussed further in Sec. 6.5.

Region filter

The results presented in Sec. 6.4.3 show that our approach can achieve an accuracy similar to previous work (cf. Table 6.1) on datasets encompassing more recent and sophisticated phishing attacks, when no filter is applied. However, this comes at the high cost in terms of performance as the mean number of regions per web page is more than 38 for both phishing and benign datasets. To reduce the number of regions, we constructed a basic filter using the region properties presented in Sec. 6.3.2 such that it discards outliers without affecting the accuracy for phishing websites. As shown in the first row of Table 6.4, applying this basic filter allows us to reduce the mean number of regions from 38.20 and 38.16 down to 17.18 and 20.61 at no loss of accuracy for phishing web pages and a drop of 0.2% for benign web pages.

Further restrictions on the basic filter can, at the cost of accuracy, further reduce the number of regions found. Table 6.4 reports the filters applied in our experiments along with the obtained measures for accuracy, mean number of regions and standard deviation. We can

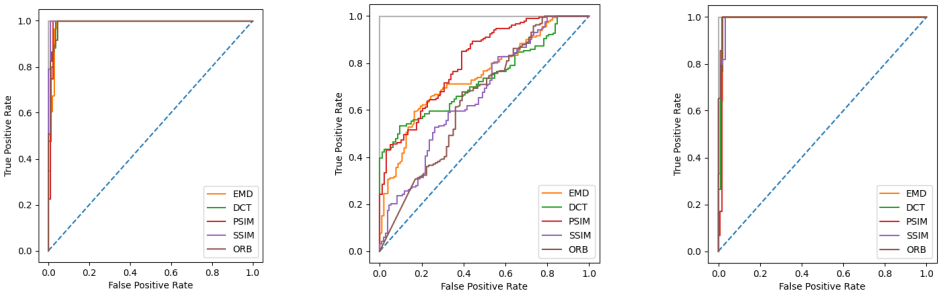
Table 6.3: Target identification accuracy for phishing and benign websites

Features	Accuracy	
	Phishing	**Benign**
Textual	86.3%	90.3%
Visual	78.6%	72.3%
Combined	99.2%	92.0%

observe that restricting the percentage of the region occupied by the most dominant color (DCP) between 9% and 75%, reduces the phishing target accuracy only by 0.6%, while reducing the mean number of regions down to 9.33 for phishing and 15.35 for benign websites. Table 6.4 also shows that the use of stricter filters can notably reduce the number of regions found while preserving a high accuracy. By applying the strictest filter, which imposes additional restrictions on all properties except for the region coordinates, the mean number of regions is 6.04 and 8.69 for phishing and benign web pages respectively. However, the accuracy of target identification for phishing web pages drops from 99.2% down to 96.9%.

Table 6.4: Performance of the various region filtering strategies. Showing strategies for combinations of the Color Count (Col. Count), dominant color percentage (DCP), Area, Height, and Coordinates (Coord.). The filter feature symbols refer to the following: '○' stands for the feature is not used, '◐' stands for the feature is used without allowing for loss in phishing accuracy, and '●' stands for the feature is used while allowing minor loss in phishing accuracy.

Filter features					Accuracy		Mean regions		S.dev of regions	
Col. Count	DCP	Area	Height	Coord.	Phishing	Benign	Phishing	Benign	Phishing	Benign
○	○	○	○	○	99.2%	92.0%	38.20	38.16	43.74	45.60
◐	◐	◐	◐	◐	99.2%	91.8%	17.18	20.61	15.03	17.96
●	◐	◐	◐	◐	98.6%	91.8%	15.17	19.48	13.91	17.64
◐	●	◐	◐	◐	98.6%	91.8%	9.33	15.35	13.45	17.09
◐	◐	◐	●	◐	98.5%	91.8%	15.31	20.14	8.81	17.82
◐	◐	◐	◐	●	98.5%	91.2%	14.62	17.91	13.16	16.64
◐	◐	●	◐	◐	98.4%	91.6%	15.52	20.27	13.88	17.81
●	●	◐	◐	◐	97.7%	91.8%	8.07	14.41	7.78	16.81
●	●	●	◐	◐	97.3%	91.8%	7.07	14.11	7.15	16.73
●	●	●	●	◐	96.9%	91.8%	6.04	8.69	6.47	8.10



(a) Using text features only **(b)** Using visual features only **(c)** Using combined features

Figure 6.9: Receiver operating characteristic (ROC) of using Earth Movers Distance (EMD), Discrete Cosine Transformation (DCT), Euclidian distance for pixel similarity (PSIM), Structured Similarity Index Measure (SSIM), and 'Oriented FAST and Rotated BRIEF' (ORB) using the Dominant Color Percentage filter.

6.4.4. Classification performance

The screenshot of the visited page is compared to the screenshot of the associated pages found during target identification to determine the legitimacy of the page (step 6 of Fig. 6.3). We evaluated the effect of the proposed target identification approach on the phishing classification of the 2,000 websites in the phishing and benign datasets using existing visual similarity-based approaches: Earth Movers Distance (EMD) [122], Discrete Cosine Transformation (DCT) [313], Euclidean distance for pixel similarity (PSIM) [366], Structural Similarity Index Measure (SSIM) [367], and Oriented FAST and Rotated BRIEF (ORB) [301]. These approaches compare two images and give a score between 0.0 and 1.0 based on their similarity. Logistic regression using 10-fold cross validation was used to compute the threshold for the classification. The Receiver Operating Characteristic (ROC) and corresponding Area Under the Curve (AUC), which shows the change of false positive rate with respect to true positive rate while varying the discrimination threshold of the classifier, are reported based on the optimal threshold for each classifier (Fig. 6.9). We used the *Dominant color percentage* filter due to its trade-off between accuracy and number of regions found (cf. Sec. 6.4.3).

The ROC obtained when visual features, textual features or their combination are used for target identification is shown in Fig. 6.9. We can observe that the use of only visual features (Fig. 6.9b) performs significantly worse than using only textual features (Fig. 6.9a): AUC of 0.5808 for visual features versus 0.9816 for textual features. The use of both text and visual features (Fig. 6.9c) is comparable to using only textual features, with an AUC of 0.9920.

Table 6.5 reports accuracy (Acc.), precision (Prec.), f_1-score and AUC for the different visual similarity-based approaches when visual and textual features are used. We can observe that using exclusively the visual features we had the worst performance across the board, followed by using the textual features alone. The combination of textual and visual features outperformed the textual features slightly, achieving a high accuracy (0.9931-0.9966), precision (0.9885-0.9955), f_1-score (0.9941-0.9977), and AUC (0.9913-0.9938).

Table 6.5: Performance of Earth Movers Distance (EMD), Discrete Cosine Transformation (DCT), Euclidean distance for pixel similarity (PSIM), Structured Similarity Index Measure (SSIM), and Oriented FAST and Rotated BRIEF (ORB) as classifiers with the DCP filter. The best performers for a metric are in bold.

Classifier	Textual Features Only				Visual Features Only				Combined			
	Acc.	Prec.	f_1	AUC	Acc.	Prec.	f_1	AUC	Acc.	Prec.	f_1	AUC
EMD [122]	0.9862	0.9803	0.9898	0.9782	0.6930	0.6741	0.8024	0.5766	0.9931	0.9892	0.9944	0.9913
DCT [313]	0.9897	**0.9863**	**0.9930**	0.9804	0.7207	0.7059	0.8224	0.5769	0.9931	0.9886	0.9941	0.9920
PSIM [366]	**0.9898**	0.9852	0.9923	**0.9858**	0.7028	0.6877	0.8114	0.5650	0.9932	0.9885	0.9941	0.9923
SSIM [367]	**0.9898**	0.9852	0.9924	0.9849	0.6851	0.6591	0.7938	0.5868	**0.9966**	**0.9955**	**0.9977**	**0.9938**
ORB [301]	0.9830	0.9734	0.9861	0.9788	**0.7384**	**0.7170**	**0.8325**	**0.5987**	0.9931	0.9894	0.9945	0.9906

The ORB classifier outperforms the other classifiers with respect to all metrics when only visual features are used. However, it is among the worst classifiers when only textual features or both textual and visual features are used to identify the associated pages. This is because ORB uses key point matching for screenshot comparison and, therefore, performs better when portions of the screenshot are almost identical and worse when key points cannot be matched correctly. In case visual features are used, associated pages are identified based on the regions in the screenshot of the visited web page and, thus, the regions in the two screenshots are nearly identical.

In case only textual features are used, DCT, PSIM and SSIM provide the best performance, with PSIM and SSIM having the highest phishing detection accuracy (0.9898). On the other hand, SSIM provides the worst results with respect to accuracy, precision and f_1-score when only visual features are used. However, the results show that SSIM is the best classifier with respect to all metrics when both textual and visual features are used for target identification.

6.5. Discussion

In this work, we showcase our approach towards unknown phishing websites detection. We evaluate the effects of target identification through textual and visual features on phishing classification which achieves an accuracy up to 99.66% (ref. Table 6.5) signaling its capability to be applied to real-world scenarios. As attackers lure users to visit a phishing website by sending a message (e.g., email, SMS), victims are often prompted to fill in their login credentials, which are then retrieved by the attacker. The tool warns users before they submit their credentials to a potential phishing website thus, supporting users' decision-making when faced with suspicious pages and enabling an integrated research line on zero-hour phishing. In the following we discuss the main takeaways from the proposed approach and its evaluation.

6.5.1. Visual features improve zero-hour phishing detection

Reverse image searching of visual features has proven to be a good augmentation to existing target identification approaches. A critical aspect of similarity-based phishing detection approaches is the correct identification of legitimate web pages that phishing attacks attempt to imitate. Current approaches often perform this by only extracting textual features from the DOM of a phishing web page and, sometimes refining search terms using properties of a language (e.g., English). Upon verifying their accuracy against more recent phishing websites, the performance of the best approach has dropped from 99.6% to 86.3% (cf. Sec 6.2.2). To detect the missed phishing attacks, we proposed an approach that relies on the visual features of the visited website for target identification. We observed that phishing web pages with dissimilar appearances, such as those illustrated in Fig. 6.10, could still be associated with their corresponding legitimate web page. This shows that the visual features can be used to identify pages that are partially copied (Fig. 6.10a) or created with poor quality (Fig. 6.10b). Table 6.3 shows that our approach can identify the targeted brand with high accuracy (99.2%). This result is comparable to state-of-the-art solutions while retaining language independence. We believe that this work provides a building block for future efforts in

(a) (b) (c) (d)

Figure 6.10: Dissimilar phishing web pages (left) and their respective legitimate web pages (right) for Facebook (a and c) and Apple (b and d).

brand identification, including a more accurate labeling of the data, which is often required in phishing studies [272].

Region properties are good discriminators for selecting regions that contain identifiable information. Our approach relies on a filter that discards regions that less likely contain identifiable information to reduce the number of regions to be reverse searched. The results show that the number of regions per web page can be significantly reduced at a minor cost to accuracy (cf. Sec. 6.4.3). In particular, *dominant color percentage* provides the best trade-off, with an accuracy loss of 0.6% and a 45% reduction of the mean number of regions for phishing websites (4th row in Table 6.4). This means that our intuition that regions with high contrast likely contains identifiable information (cf. Sec. 6.3.2) holds and such heuristic can be used to reduce reverse image searches by excluding irrelevant regions. The considered region properties used for filtering do not account for the properties of the entire screenshot. For example, a low number of unique colors in a region could be normal if only a few colors are used in the web page. Also, our approach heavily relies on reverse image searching, which may encounter difficulties due to real-world constraints (e.g., low internet speed or internet data cap). Further improvements on the filtering, e.g. using OCR or language-specific properties, could help mitigate these issues [4].

Visual features have a positive impact on classification. From the phishing classification analysis, it emerges that using visual features for target identification in addition to textual features has a positive effect on all reported metrics (ref. Table 6.5). The highest accuracy when only textual features are used and when both feature types are used is achieved by SSIM, with an improvement from 98.98% to 99.66%. Whereas the improvement is small in absolute terms, it represents a $\approx$ 70% reduction in error rate. This is in itself an important step forward as with high numbers of websites to check even small error rates can get in the way of the usability of a decision support approach based on webpage analysis techniques. An interesting direction for future work is to test our approach to target identification using convolutional neural networks, shown to outperform current image similarity classifiers such as EMD, DCT, ORB and others [3, 190, 210, 213].

[4]The latest version of the anti-phishing tool uses a random forest classifier to retain only regions with relevant information based on the regions' features, thus, avoiding using the heuristic filter strategies. The tool is available at: `https://github.com/paolokoelio/zerohour-decisionsupport-phishing`

6.5.2. Research directions for user experimentation

As zero-hour phishing detection methods can generate false positives, human intervention is still needed in the decision making process. This, however, places additional burden on the user. To this end, research should assess the best ways to avoid too much strain on the user while keeping them safe. Our visual similarity-based phishing detection tool enables this line of research. The tool is packaged into a usable and upgradable browser extension and a web API. This allows a fast deployment of experiments with a scalable number of participants to investigate research gaps in this area. We identify three main research directions that could be supported, experimentally, by the proposed tool:

Assessing user aids supporting decision-making on website legitimacy. Thanks to prior research in usable security, passive indicators have been replaced with blocking warnings. Nonetheless, new experiments can shed light on the gaps not filled by active warnings, such as the circumstances of warning triggering. Our tool can be used to evaluate (types of) warnings in the context of different website categories, such as e-commerce, social media or banking. Similarly, different implementations of nudges, such as dynamic notifications or timed blocking of the "Submit" button, can be tested in various circumstances. For example, experiments can be set up within an organization's embedded phishing training, thus allowing warning efficacy to be tested in an ecologically valid setting.

Evaluating user trust in a detection system's risk advice. The efficacy of decision-support systems depends on the balance between system's capabilities and users trust [70]. Our tool can help investigating the calibration between the perceived trust and the tool's risk advice by dynamically customizing the warning contents. For example, effects of user calibration on the final decision can be measured by presenting further details on where, how and when a warning has been generated or by displaying the tool's detection statistics. Research on indicator proxies for the inner processes of the tool, such as progress bars or status indicators, has the potential to steer user perceptions and, eventually, improve user choices. Experiments can benefit from the dynamic interaction of the plug-in and the underlying detection logic where, for example, experiment designs may vary the content and placement of status indicators in the browser UI at detection run-time.

Exploring new risk communication methods by keeping track of past decisions and associated risks. Whereas visual similarity-based detection tools are able to detect zero-hour attacks, they have typically long runtimes, which can significantly affect a user's reliance on such tools. Our implementation takes an original approach to this problem: instead of blocking users before the detection is complete (as done by, e.g., Microsoft SafeLink), users are notified retrospectively of the past phishing encounter and, thus, can remediate 'bad' decisions by changing their credentials. While a similar approach has been successfully applied against credential stuffing attacks, it is unclear if this concept is effective in a near real-time setting. Our tool enables further research in this direction, for example, user studies on the efficacy of retrospective notifications to reduce attack success rates.

6.5.3. Limitations and threats to validity

The search engine is a critical third-party component of our approach, which relies on the assumption that search engines typically show benign results first and that the best match appears among the top results. The benefit of using a search engine is that no local database of all benign websites has to be maintained. A desired behavior is, therefore, that the results from a search query change over time so that new websites are also protected without relying on client-side information. However, this changing behavior implies that a search engine can return different results for the same query over time. This highlights that, especially for pages where the brand name cannot be extracted from the DOM and a set of generic terms is extracted instead, a variance in accuracy can occur. Our approach partially addresses this issue by accounting for visual features extracted from the screenshot of the web page, making it more robust when brand names do not occur in the text.

To ensure generalizability of our approach, the datasets were constructed in such a way they are representative of the Internet at large. The phishing dataset used during testing was representative of recent sophisticated phishing attacks as reported by OpenPhish [258] and APWG [28]. The benign data is based upon an open directory so that it is not exclusively composed of websites with large traffic, making it more representative of the websites users typically visit. However, global trends do not necessarily translate to local trends. For example, a dataset containing phishing Facebook pages may not be representative for countries where local social network are largely used (e.g., `vk.com` or `qq.com`). This is unavoidable for any phishing study that does not target specific demographics. It is worth noting that all entries in the dataset were manually verified to mitigate inconsistencies arising from automatic labeling.

A threat to internal validity is the calibration of the region filters used to reduce the number of regions in a screenshot (cf. Sec. 6.4.3). The use of different properties (and values) could have achieved a different reduction in the number of regions or accuracy loss. Here the filters were defined experimentally; this does not guarantee that they can achieve the same results with different datasets.

Finally, the detection time is largely bounded by the requests to and replies from the search engine which can compromise the final usability. We have not evaluated the detection speed at runtime as the scope of our efforts is delimited to the improvement of detection performance and enabling experimentation with decision support systems. Still, to mitigate long waiting times, the system will alert the user retrospectively in case the user navigates away from the not-yet-detected phishing page. Future work can investigate time performance improvements, such as optimizing a caching system to reduce the overall number of requests.

6.6. Conclusion

We presented a novel approach for zero-hour phishing detection that uses both textual and visual features of a potential phishing page as search terms to identify the legitimate mimicked website. By relying on visual features extracted from a screenshot of the web page, our method is robust against image-only brand embeddings. An evaluation of our approach shows that the use of visual features, in addition to textual features, allows achieving an accuracy of 99.2% for target identification and of 99.66% (vs. 98,88% text-only) for phishing classification on a dataset comprising phishing attacks from 2019. The improvement of 67% reduction in miss-classification rate represents an important step forward in phishing detection as even few false positives can compromise user trust in a decision support system, degrading its effectiveness altogether. By integrating phishing detection and human-computer interaction ingredients together, we propose a new experimental tool to evaluate, characterize, and refine the interaction between zero-hour phishing decision support, and the final user. Future developments can improve the runtime performances of the tool as well as evaluate its usability.

In Chapter 5, we have seen that an advanced tailoring of phishing artefacts makes phishing attacks particularly effective. The mitigation method proposed in this chapter aims to warn and protect users from revealing credentials to a phishing website, including phishing web pages tailored to virtually any organization or user base, as the tool is agnostic of the targeted base. This technological mitigation strategy contributes to answer the second part of RQIII. In the next chapter (Chapter 7), we follow up with the second part of RQIII by observing how the untapped potential of crowd-sourced detection and reporting can provide an organizational mitigation against advanced phishing attacks, such as tailored phishing.

7

Phishing reporting as an untapped defense strategy

Albeit technological solutions can support users with certain classes of attacks, such as phishing websites in Chapter 6, existing prevention and detection countermeasures may be not enough to thwart advanced phishing attacks, such as spear and tailored phishing. Such attacks are difficult to detect automatically due to the large variance in attack characteristics, including chosen pretexts, multiple attack vectors, and the 'dilution' of specific attack signatures across multiple stages. Artifacts are meticulously crafted to fit targets' context and to fly under-the-radar, for example, by employing legitimate or vanilla websites, and targeting small numbers of recipients. Anomalies in the communication still exist, but they are hard to formalize and cannot be captured automatically by a single technological solution. In this chapter, we propose an organizational mitigation approach to address the limitations of existing countermeasures (see RQIII). Our proposition is a new course of action to exploit human detection capabilities as a potential basis of automated response strategies. Preliminary results unveil users' mental models for phishing detection and reporting as a way to improve the phishing reporting process and lower the number of victims. A real word case study is provided to showcase the feasibility of our proposal.

This chapter is originally published as P. Burda, L. Allodi, and N. Zannone, "Don't Forget the Human: a Crowdsourced Approach to Automate Response and Containment Against Spear Phishing Attacks", In *IEEE European Symposium on Security and Privacy Workshops* (EuroS&PW), IEEE, 2020, pp. 471–476

7.1. Introduction

Chapters 3 and 4 showed us that phishing techniques provide the advantage that the 'human vulnerabilities' they attack cannot be easily fixed, and are approximately the same across targets. For this reason, such attacks are evolving rapidly into new, more sophisticated attack scenarios to overcome available countermeasures. Being this a technologically cheap yet powerful exploitation technique, it has become the preferred method employed by attackers to compromise systems and exfiltrate data from individuals as well as large targeted organizations [350, 352].

The latter in particular represents an increasingly worrisome trend of sophisticated, highly-targeted social engineering attacks [350, 352] against which common countermeasures aiming at 'general' phishing are ineffective [15]. These attacks are 'tailored' against specific organizations or groups of people, and differ significantly from generalist attacks. For example, by means of multiple iterations and reconnaissance, an attacker can tailor social engineering artifacts that can be extremely effective on their targets [47]. Cognitive attacks aimed at persuading victims in executing an action are diluted in multiple interactions exploiting the communication methods and language the organization is used to, making them hard if not impossible to detect by traditional means.

Following on Chapter 2, state of the art counter-measures integrate training of employees, advanced software and security operation centers. However, existing countermeasures are lagging behind the expansion of sophisticated phishing attacks, first of all, spear-phishing [149]. Attack features like content and links are extremely variable, hindering the majority of detection attempts or generating too many false positives [77]. Further, the resemblance of these attacks to regular communication make training and awareness campaigns largely ineffective to 'immunize' a large fraction of the victim pool [62]; anomalies in the communication still exist (e.g., unusual references to internal processes in an organization), but these are hard to formalize and cannot be captured automatically by a single technological solution. Therefore, organizations often rely on response teams, like SOCs and CERTs, as the last line of defense. However, current containment procedures based on after-the-fact analyses are too slow to match the high velocity at which spear-like phishing campaigns are known to affect their targets [188].

In this chapter, we propose a way ahead to respond and contain advanced spear-like phishing attacks in organizational settings (see RQIII). Our proposition aims at leveraging the natural 'immunity' of (some) human subjects in an organization (e.g., senior employees with a deep knowledge of the 'normal' processes within the organization and a natural ability to detect 'anomalies') to mitigate and contain the attack against the organization as a whole. At the core of the proposed solution is a more efficient *phishing reporting process* based on cognitive mental models of individuals better predisposed to detect complex attacks. An improved reporting process, potentially merged with automated response procedures, can allow to speed up the containment of an attack, thus lowering the number of victims.

The remainder of the chapter is structured as follows. Section 7.2 highlights the identified research gaps. Section 7.3 provides a description of the proposed solution, a motivating example, and our research plan. Section 7.4 presents preliminary results.

7.2. Open problem

In Chapter 2, we described how existing countermeasures, while somewhat effective against generic phishing, are inadequate against its advanced variants, such as spear- and tailored phishing. Training and threat awareness are unable to make subjects immune to sophisticated attacks, leaving a large fraction vulnerable to them. Similarly, detection techniques are not able to cope with the large variance in spear-phishing attacks, including chosen pretexts, single vs. multi-stage attacks, and the 'dilution' of specific attack signatures across multiple communications or phases of the attack. Anomalies in the communication processes and protocols characterizing a specific organization may represent an unexplored venue for research, but these processes and protocols are hard to formalize; as a result, a general anomaly-detection solution for spear-phishing applicable to any organizational settings is not on the horizon. The fundamental problem is that the specific characteristics of spear-phishing attacks (multistage processes, tailored artifacts, yet-to-be-seen malicious URIs, etc.), make current defensive techniques inadequate and ineffective.

The consequences are well signaled by industry reports which point at phishing attacks to be the most prevalent attack and source of compromise for most organizations [350]. Therefore, new approaches are necessary to cope with spear-phishing attacks for which both prevention and detection are fundamentally unsuited as the prevalent defensive barrier.

7.3. Proposed solution

Due to the foundational differences between phishing and spear-phishing, prevention and detection techniques may be grossly inadequate to tackle the problem. However, we believe there is an important gap in the *response* phase that could provide large benefits to organizational security: human reporting is an untapped resource that could provide readily available risk indicators for suspicious emails, and lead to fast attack response and containment. This requires increasing the quality of phishing reports and, possibly, automating a risk-based containment phase to promptly react to a reported attack. Preliminary evidence (see Sec. 7.4) suggests that some users are naturally predisposed to identify anomalies between the communication processes employed by spear-phishing attacks and the 'normal' ones employed by an organization. However, only a few users typically report phishing emails, and the rationale and incentives behind this are still unexplored in the scientific literature. Once deconstructed, the *mental models* behind phishing reporting could be employed to increase reporting incidence, speed, and to build reputation-based methods (like in [212]) to assign risk-scores to specific (likely) attacks. Moreover, the few users that report suspicious emails do this *immediately* when they receive them, providing a timely information source to employ for response. Yet, this is currently untapped. Based on user reports, further developments of the proposed solution can leverage a portfolio of automated response strategies, such as issuing warnings to other users, automated URI blacklisting, AV signatures generation, centralized filters, etc., to protect the large fraction of users that have not *yet* fallen for the attack, but that most likely would if no response is put in place.

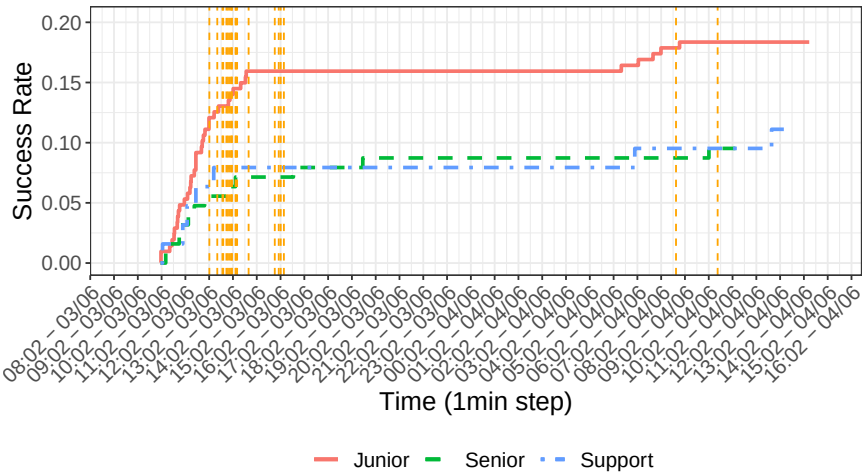


Figure 7.1: Success rate over time per user category. The vertical lines indicate the time our emails were reported.

7.3.1. Motivating example

Previous findings in the literature [161, 192] already showed that (spear-)phishing attacks trigger victim responses very quickly. However, it is unclear what is the potential of reporting mechanisms to provide timely information on the attack. To provide a first evaluation of this, we look at the tailored phishing campaign we ran against our university for internal measurement purposes in Chapter 5, in collaboration with the security department of the university. The campaign pretext asked users to participate in an HR process to collect vacation hours, a process that is *not* employed at our university. Figure 7.1 reports the rate of users falling for the attack (i.e., that would have submitted their credentials if this was a real attack) by user category. Notice that Junior employees (PhD students and postdocs) are those that fall for the attack the quickest and by the largest fraction. By contrast, senior scientific staff and support staff are much less vulnerable overall, further supporting the intuition that expertise on internal processes may be a decisive factor in the successful distinction of a spear-phishing attack of this type. Regarding the velocity of the attack, approximately 75% of employees that fell for the attack did so in the first four hours since the start of the campaign. Interestingly, 23 employees detected and reported the attack to the IT department, many of which when the campaign was at its peak (vertical lines in Figure 7.1). Our intuition is that a containment action during the (first few) incoming reports can eventually reduce the victimization rate by 25 to 40 percent by a swift blocking procedure.

7.3.2. Improving the reporting process

The core of our proposition to make the security process more efficient is to leverage the already present human detection capabilities of 'phishing champions' to improve the reporting process. From an organizational point of view, companies can benefit from employees

that notify the IT departments in case of abuse. The efficiency of this reporting process, however, depends on the number and the quality of such notifications. We hypothesize that, among the employees of an organization, there are some that are particularly good at detecting phishing, further down referred to as 'phishing champions'. However, not all of those are keen to report suspicious emails. We are interested in fostering reporting only from champions to keep the noise/signal ratio to favorable levels.

The proposed course of action is to devise methods able to identify phishing champions, and to do so, we first need to understand what are their characteristics. We look for characteristics that may correlate with higher detection skills (e.g., experience) and reporting eagerness (e.g., sense of responsibility). We can collect this qualitative data by means of structured interviews of an organization's employees that have reported phishing attacks in the past and potentially those who have not reported but detected them. By employing qualitative analysis methods to analyze the interviews' answers, we can extract actionable topics and reconstruct the *mental models* users follow when deciding whether to report a phishing email. Mental models go beyond simple schemata of highly regular and routine situation (like trivial phishing) and can better represent new situations through the use of generic knowledge of space, time, causality, and human intentionality [173].

Based on these mental models, we can design a diagnostic test to systematically identify 'champions', including those that are not keen or aware of the reporting process, and develop risk-based metrics to evaluate the uncertainty around the report. These metrics can be based both on past reporting activities of the employee, as well as specific characteristics of the reported artifact.

7.4. Preliminary results

As a first effort towards the identification of phishing champions, we interviewed the employees that reported our phishing attempts to the IT department during the phishing campaign of Chapter 5, as shown in Figure 7.1. Specifically, we were able to interview 12 out of the 23 employees that reported our phishing email. Following a one-page interview guide [46], we first asked high level questions on detection and reporting. Then, we invited the interviewee to retrieve and read the e-mail they received (if needed, we provided a printed copy) and asked detailed questions with the specimen at hand. The interviews investigated how does the interviewee:

1. detect phishing emails in general and the specific email they received
2. decide to report phishing emails in general and decided to report the specific email

The semi-structured interviews were recorded and transcribed. The interviewees' answers were coded using a card-sorting technique to derive mental models reflecting the decision process of the respondents. The more similar users' thoughts are between the general case and the specific attack, the more 'mature' the mental model can be considered to be, as it characterizes users that can abstract their reasoning away from examples without loss of information. By contrast, mental models that are much more detailed when example-driven than in the general case suggest a less mature rationale whereby users cannot abstract away from the example.

7.4.1. Results

Figure 7.2 presents the results concerning the detection of phishing emails in general (left) and the detection of the phishing email sent within our campaign (right). We include subjects' characteristics, namely the seniority (Junior, Senior and Support) and department they work at (WIN - Mathematics and Computer Science, and SWT - Chemical Engineering). Arcs are labeled with numbers identifying the interviewees which followed that arc in their reasoning. For example, in Figure 7.2b, ID 7 is a senior employee whose detection method can be reconstructed following the arcs labeled with ID 7, starting from the root: 1) the content's semantics does not match her previous experience, 2) an unfamiliar signature is present, and 3) there is a link with a wrong domain name. We can observe that, for detection, the mental model derived from the concrete example (Figure 7.2b) largely matches the mental model derived from the general case (Figure 7.2a). In contrast, there is not a clear overlap between Figures 7.3a and 7.3b, suggesting a less developed mental model for reporting.

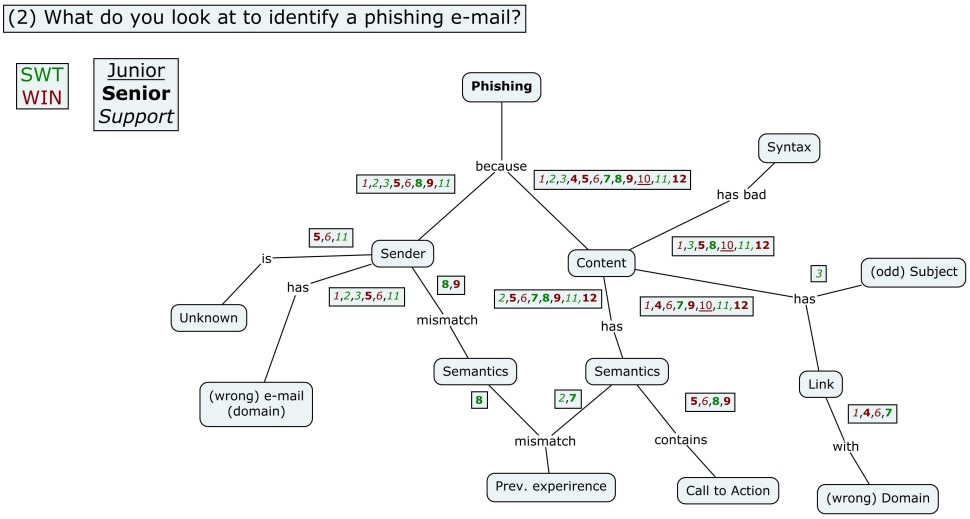
How do our interviewees detect phishing attempts

In the interview, we first asked users how do they detect *general* phishing attempts, and why did they detect our specific phishing email. Answers to the first question prevalently refer checking if the content's syntax is correct, if the semantics 'makes sense' to the user and if the sender's email domain is correct. When prompted with the email from our campaign, the reasons why our respondents detected the phishing email as such largely overlap with the answer they gave us in the general case. For example, Respondent 3 states "*It's a follow up? Strange request. Why is there a link? Strange email…*", highlighting the 'anomalous' nature of the email w.r.t. what she is used to receive from the department. Similarly, Respondent 8 answered: "*It's a bit weird for UNI to ask me my holiday hours, I already have them on the [HR's portal]. Also the domain is not good [..]. It's asking for specific action which does not apply to me.*", further highlighting the email inconsistencies also remarked by many for the detection of a 'general' phishing email. On this same line, Respondent 9 answers: "*The sender does not match the topic semantics, it's like you have a painting from Rembrandt and suddenly you have an iPhone there*".

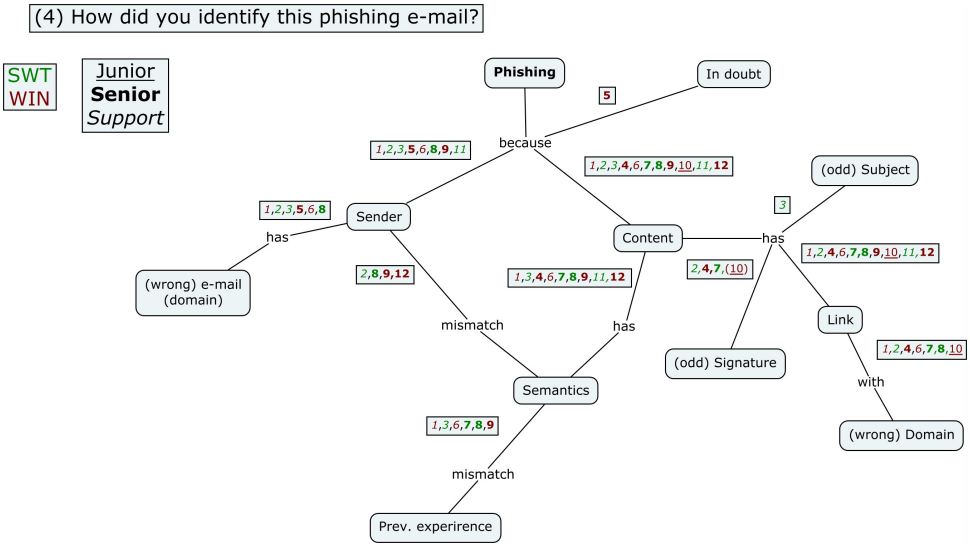
In all, we find our respondents were very consistent in their rationale to classify a phishing email as such, whether this is a 'general' or hypothetical phishing email, or a concrete example with which they are already familiar.

Why do our interviewees decide to report phishing emails

When asked if they usually report phishing attempts, the majority of respondents answered that they do not report phishing on their personal (email) accounts, and reporting at work happens more as an exception. As for the question why they do report phishing emails, answers are very general and a clear rationale does not emerge. For example, Respondent 1 answered: "*[I report when] I'm in doubt it could be a legit email*", highlighting uncertainty as a key element in their decision to act on the attack. On a similar line, Respondent 6 states: "*I report if I know the sender, e.g., my bank or my organization*", suggesting that only emails that are 'relevant' to the user's context will be reported by this user. However, a clear-cut reason



(a) General case

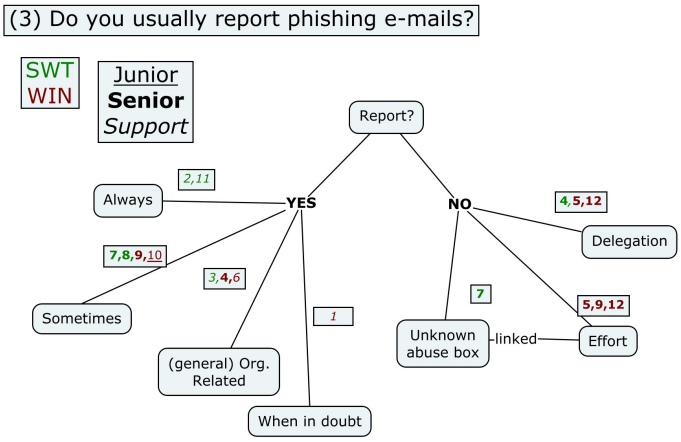


(b) Concrete case

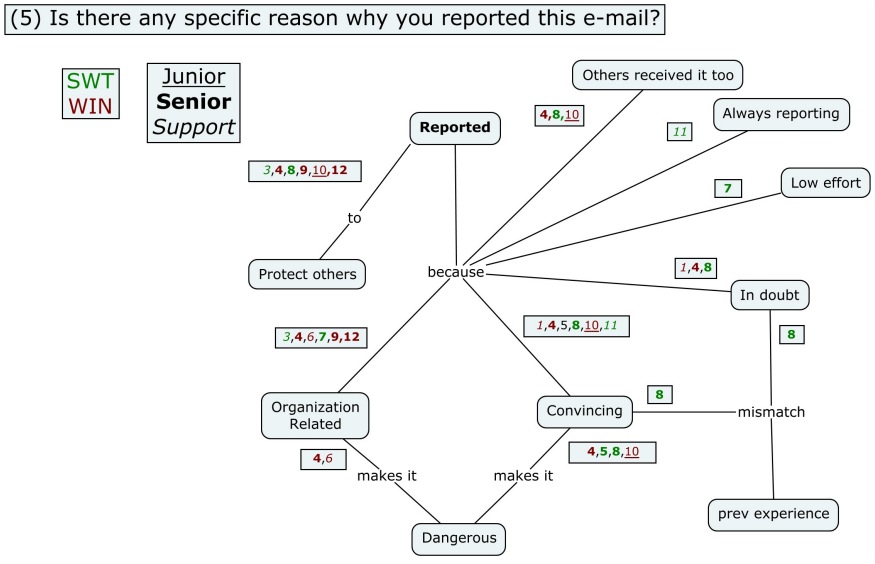
Figure 7.2: Mental model of phishing detection

to discern between 'general' phishing emails that our respondents will report, and those they won't did not emerge.

By contrast, the reasons to report our specific phishing email appear to be much more structured and detailed, and include reasons relating to safeguarding less-aware colleagues and the perceived sophistication of the attack. For instance, Respondent 3 states she reported the email because "*it pretends to be from UNI, to protect others*"; similarly, Respondent 6



(a) General case



(b) Concrete case

Figure 7.3: Mental model of phishing reporting

says he reported the email "*Because it's posing as UNI, my organization should know about it*". In sharp contrast with the detection case, the respondent's mental models for phishing reporting appear to be much less developed, structured, and consistent, suggesting a strong imbalance in user prowess between detection and reporting activities.

7.4.2. Discussion

Mental models showed the respondents' inability to generalize the rationale for notifying a suspicious email, thus providing insights on where improvements of security processes can be made. For example, answers highlight that the reporting procedure is ill-perceived in terms of effort and liability ("Effort" and "Delegation"[1] in Figure 7.3a). Such models can also shed light on the underlying factors for reporting, such as a higher sense of responsibility and threat awareness ("Protect others" and "Dangerous" in Figure 7.3b).

The results, however, may be influenced by the specific type of organization where the study was carried out. Other domains, like financial or industrial, may lead to different outcomes both in terms of reporting rates and reasons to report. More studies are needed to generalize our results.

From our preliminary evaluation, a more thorough and rigorous investigation could shed additional light on *users' rationale to report phishing and factors that influence their decisions*. Future work could tackle new research in this direction by evaluating the training and reporting problem, for example by investigating whether an efficient phishing *reporting* process can aid the protection of users that fail to detect the phishing email as such.

7.5. Conclusion

In this chapter, we argued the urgency for new paradigms to counter advanced phishing attacks. In particular, we proposed a new course of action to address the limitations of existing countermeasures against this class of attacks by exploiting human detection capabilities as the basis for improved reporting procedures. We are guided by the intuition that a sufficiently high proportion of phishing immune individuals can help those that are not and aid the resilience of the organization as a whole. Preliminary results show how to measure users' mental processes (i.e., users' rationale and decision making) as a way forward to improve the phishing reporting process altogether. We promote this idea using a real world example and provide directions on how to make human reports actionable.

Our findings support an organizational mitigation strategy against tailored phishing attacks by promoting a research line that investigates user reporting behaviors. Together with Chapters 5 and 6, this contributes to answer RQIII. The preliminary results of this chapter suggest that exploring users' rationale and factors that influence user behavior in relation to phishing reporting can shed light on how to mitigate the impact of tailored phishing attacks. In the following part (Chapters 8 and 9), we explore this research line by investigating RQIV.

[1]By "Delegation", the respondent assumed it is someone else's duty to deal with security incidents.

III

Why do people report phishing

8

Collective phishing defense mechanisms at a small IT company

Extant research has primarily investigated tailored phishing campaigns in the context of large enterprises. On the same line, Chapter 7 introduced the idea of instrumenting the natural ability of employees to detect 'anomalies' to report phishing attacks at organizations. Company size, composition, and resource availability (such as, security experts or a phishing response team handling incidents) play an important role on the effect of such dynamics. However, whether the same also applies to small and medium-sized enterprises (SMEs), which typically do not have those resources, is unclear. On the other hand, studying SME security is hard as they generally have no expertise in-house to run the required experiments. This chapter fills this gap by investigating the response to a tailored phishing campaign of employees of a small IT company. To this end, we conducted a field experiment targeting 30 employees at a small research and development company in the Netherlands. Subsequently, we interviewed nine employees with mainly technical background to understand the cognitive processes underlying the detection and response of our phishing campaign, as well as the group defense mechanisms at the SME (see RQIV). Our findings show that the collective defense mechanism enabled a surprisingly prompt response and containment of the attack, possibly, due to the network dynamics of a small company.

This chapter is originally published as P. Burda, A. Altawekji, L. Allodi, and N. Zannone, "The Peculiar Case of Tailored Phishing against SMEs: Detection and Collective Defense Mechanisms at a Small IT Company", In *IEEE European Symposium on Security and Privacy Workshops* (EuroS&PW), IEEE, 2023, pp. 232-243

8.1. Introduction

Throughout Part II, we have seen that tailored phishing can lead to a high impact at organizations with relatively low effort from the attacker. In this chapter, however, we observe that the extant literature has primarily focused on large companies [63, 161, 274, 317, 371]. The reasons are multi-fold: large companies have a higher number of employees, which is desirable for experimentation; have IT and oftentimes security teams that are interested in the outcomes of the experiments/measurements or operationalize part of those (e.g., an internal phishing response team); are typically large enough that budgeting and employee time is not an issue when designing and deploying the experiment. On the other hand, a very large portion of the economic substrata of developed countries is constituted by small-medium enterprises (SMEs). This kind of enterprise has nowhere near the resources and technical capabilities of large enterprises [109]. Critically, SMEs are often part of the *supply chain* of large organizations, meaning that an attack on one can have rippling effects on the whole chain [107]. Yet, few insights exist on the effectiveness of tailored phishing attacks in SME contexts, and which factors, if any, can play a role in SME resilience against these.

In this chapter, we fill this gap by investigating the effectiveness of a tailored phishing campaign against a small (35 employees) company in EU (hereafter referred to as *CompanyX*[1]), and evaluating employees' response to it by means of semi-structured interviews with nine participants. We first run a field experiment demonstrating how an attacker can leverage publicly available data in a tailored phishing attack, evaluate the OSINT attack surface of an SME, and evaluate its effectiveness as a phishing campaign. Interestingly (and perhaps surprisingly), we observe the engagement of a strong community reaction around the attack whereby employees take immediate action to protect each other from the attack. To investigate this in more depth, we run a set of interviews with company employees and uncover the cognitive and behavioral aspects that ultimately triggered this collective response mechanism.

Our findings show that expectation mismatch – the detection of an inconsistent pattern – was the primary method for detecting our phishing campaign, possibly due to the limited size of the company where 'everyone knows everyone'. The tailored nature of the attack provoked some employees to take action and alert the rest of the group, yielding a defense reaction surprisingly faster than what is expected in larger organizations. Our interviews show that the collective defense might be linked to the network effects of a small organization that mitigated the bystander effect, which on the other hand is often observed in large organizations.

The remainder of the chapter is structured as follows. Section 8.2 provides background information and discusses related work. Section 8.3 introduces our research questions and the methodology employed to address them. Section 8.4 presents the results, and Section 8.5 discusses lessons learned and limitations of our work.

[1]The actual name of the company cannot be disclosed due to confidentiality reasons.

8.2. Background and Related work

Background

As seen in Chapter 2, tailored or spear-phishing variants, i.e., phishing emails exploiting target-related information for attack engineering, pose significant threat to their targets due to their overall higher efficacy over 'regular' phishing [150]. Their success, however, depends on an attacker's ability to gather detailed information about their targets. This can be achieved by employing open-source intelligence (OSINT) techniques to gather publicly available information, such as personal and professional information on social media or public records [37, 103]. However, as shown in Chapter 1, there is still little understanding of what kind of target-related information might influence the effectiveness of a phishing attack.

Yet, OSINT-enabled phishing attacks can have a significant impact on organizations. Small and medium-sized enterprises (SMEs) are particularly at risk due to their potentially limited resources for defending against such attacks given, e.g., smaller IT departments and limited budget [264]. Whereas research and practice devoted significant efforts to defend against phishing in organizations, such as detection software and training programs, current countermeasures might not be well-suited against tailored attacks and, more so, in SMEs [56]. Therefore, the last line of defense lies with the employees that can detect and potentially react to such attacks [195]. Proper actions employees can carry out are, for example, reporting phishing to the IT administrator or other colleagues and, if fast enough, the attack impact can be mitigated or even nullified [56, 196]. However, as organizational culture, social dynamics, and human interactions within small companies may differ significantly from those in larger organizations, which are typically the focus of phishing research, it is unclear how employees at SMEs react to said attacks and what can influence their actions.

Related work

Previous experiments on the effectiveness of tailored and 'spear'-like phishing attacks investigated the usage of information available on social media [161, 317], age and life domains (private context and different pretexts) [256], persuasion techniques [61, 63] or their combinations [291, 345]. For instance, Vahdad [345] surveyed target-related variables obtainable with OSINT that can affect phishing success and proposed a framework to instrument target-related OSINT-enabled variables to carry out phishing attacks. Potentially effective variables range from target's social network to likes and interests, including residence- and work-related information. Many studies investigate phishing susceptibility in organizational contexts [61, 63, 161, 274, 317, 371]. For example, Williams et al. [371] investigate the influence of information overload within the work environment, and Petric and Roer [274] how the organization size affects susceptibility. Notably, one study [317] investigated the susceptibility of a large organization to targeted attacks on LinkedIn and collected employees' views on threats, challenges, and overall awareness of phishing attacks on social media. Extant research on phishing reporting highlighted the importance of user notifications [196], investigated why reporting is scarce [179, 195], the rationale to report [56] and factors behind the intention to report [222]. The study in Chapter 7, explores user mental models for phishing detection and reporting in the aftermath of a simulated phishing campaign at a

mid-sized university.

Whereas studies on reporting mainly conducted surveys investigating what influences intention to report [179, 195, 222], our work measures, through one-to-one interviews, a set of user reactions to the attack in terms of explicit thought and emotion recollections that cover aspects such as attack detection, message tailoring and performed actions. We leveraged two target-related variables (**Place of Residence** and **Years in Current Company**), which we use to adapt the phishing pretext in our field experiment, since these variables can be effective in increasing the success rate of a (non-tailored) phishing attack [345]. For instance, knowledge of a target's place of residence can be highly effective in enhancing the believability of a phishing email [256]. Also, including information regarding the person's work experience increases the personalization of the message [341]: research and practice show that basic email personalization can considerably increase email opening rates [54, 79]. These variables closely relate to the target *personal* and *work-related* parameters discussed in Chapter 3. The variables are used here to craft the attack stimuli to align with the targets' context.

Similarly to [317], we interview company employees to examine their attitudes to the attack; however, our work concerns tailored attacks via email against a small enterprise, and we explore the possible factors that contribute to employees' attack detection, their reaction to the attack and undertaken actions. Indeed, existing studies have mainly focused on large organizations leaving the context of small enterprises largely unexplored. No other known work evaluated the effects of targetization in a tailored phishing attack contextualized in a small company nor explored in depth the employees' reactions to said attacks.

Research gap and contribution

The effects of phishing attacks against small companies are still unexplored. To date, it is still unclear how employees at smaller companies respond to tailored phishing attacks and which factors influence their actions. Our study attempts to address this gap by conducting a field experiment to study the effectiveness of a tailored phishing campaign targeting an SME. Finally, given the growing interest in using crowd-sourcing for reporting phishing and the lack of comprehensive research on why employees report tailored phishing attacks (including their reactions and reasoning on reporting), there is a need for a better understanding of phishing reporting in the case of tailored attacks, particularly in the context of small enterprises. Our study fills this gap by examining the reactions of the employees targeted by our phishing campaign using interviews, thereby providing insights into how SME employees react to and reason upon tailored attacks, which is critical to improving their overall security posture.

8.3. Methodology

Our study aims to study the response of employees of small organizations to a tailored phishing attacks (see RQIV). To be able to characterize employee reactions and perceptions of the attack, we first need to investigate the effect of the tailored campaign. Therefore, we structure this study along the following research questions:

RQ1: What is the effect of a tailored phishing attack on employees at a small enterprise in terms of success rate?

RQ2: How do employees of a small enterprise react and perceive the attack?

8.3.1. Methodology overview

This study uses a mixed-method approach: 1) a field experiment to assess the effects of a phishing attack tailored against a technology-driven SME and its employees (RQ1), and 2) interviews to gain insight into the motivations, emotions, and other factors influencing the employees' detection and response processes when dealing with tailored phishing attacks (RQ2).

Specifically, we conducted a controlled phishing experiment on 30 employees of an SME, where a tailored and a non-tailored phishing email were sent to their employees. The success rate of the phishing attack was measured by the number of employees who submitted data on the forged landing page. Thus, the field experiment aims to provide quantitative data on the effectiveness of tailored phishing attacks. Subsequently, nine employees from the SME were interviewed. The interviews were semi-structured, and the questions were designed to collect qualitative data about the participants' experience, thoughts, and feelings about the phishing attack, and other factors that influenced their detection and response process.

8.3.2. Subject selection

The present study was conducted at a small Dutch IT company with approximately 35 employees, with a diverse mix of international personnel. As a research and development firm, the company has a mainly technical workforce, a handful of business associates, and limited supporting staff. Our sample for this study consisted of 35 employees initially provided by the company; five of them were subsequently excluded due to insufficient online information needed to target them in our campaign, resulting in a final sample size of 30 participants.

8.3.3. Phishing field experiment

Design

To investigate the effectiveness of tailored phishing attacks against SMEs, we selected two variables as treatment, namely ***Place of Residence*** and ***Years in Current Company***, which are categorized as personal and professional variables, respectively [345]. These variables were chosen because they are relatively easy to obtain with OSINT and demonstrate the minimal effort required to craft a tailored phishing email, as well as being relevant to employees both personally and professionally (cf. Sec. 8.2). We used the two variables to create two experimental conditions: the use of *both* or *none* variables. This decision was made to account for the limited sample size of 30 employees. The employees were randomly divided into two groups of 15 each, with one group receiving a non-personalized baseline email (i.e., no treatment) and the other group receiving a personalized email (i.e., the treatment). The treatment is a modifier to the baseline email. This distinction between the two groups allowed us to observe how targets responded to tailored and non-tailored phishing emails and to determine

the relative increase in effectiveness of adding personalization to the attack. We measure the number of visits at the URL payload (clicks) and data submissions on the forged landing page. The attack success of the phishing campaign is measured in terms of submissions on the landing page.

Preparation

Initial reconnaissance was conducted using OSINT by collecting data from public sources, such as LinkedIn and Facebook, to gather information on the variables of ***Place of Residence*** and ***Years in Current Company***. These variables were later used to tailor the phishing email to the targeted population. An analysis of the typical emailing style and signatures within the target organization was conducted to make the interaction appear as credible as possible. This was achieved by sending an email regarding an internship listed on the company's website, as this approach mimics a tactic that an external attacker may use to identify internal communication practices.

The next step was the pretext building. To enhance the authenticity of the email, we agreed with our contacts within the company to impersonate the CEO. The pretext of the email, which is presented in Appendix C.1, was crafted with the collaboration of the company. The pretext concerned recently abolished Covid-19 measures and included a (fake) voucher link for a gift from the CEO that could be used at a nearby activity. The phishing email included an opening with the name of the receiver and a link that was consistent with the pretext. The language used in the email text was English, which is the standard language for official communication in the company, reflecting its international workforce.

The participants were randomly divided into two groups of 15 each; one group was assigned to a baseline email not tailored to them (i.e., no treatments) except for their name at the beginning of the email (see Appendix C.1.1); the other group was assigned to the tailored email (i.e., the treatment) with information about their place of residence and years of employment with the current company (e.g., see Appendix C.1.1).

We crafted a landing page that prompted users to submit their credentials. The company relies on Google as the email service provider; therefore, the landing page includes a Google-like login form and the logo of the company. Upon submission, targets are redirected to a debriefing web page that informs the participants about the experiment; we refer to Appendix C.2 for more details.

The final preparatory step before the execution of the experiment was to register a domain suitable to the attack scenario: we registered a domain for both sending the emails and hosting the landing page. The domain name was very similar to the actual domain name of CompanyX and differed by one letter and used the suffix `.nl` instead of `.org`.

Execution

We automated the experiment using the phishing simulation toolkit presented in [281], which extends GoPhish (`https://getgophish.com`) to enable tailored phishing campaigns at scale. The tool was supplied with the collected data, including the employee name, ***Place of Residence***, and ***Years in Current Company***.

Before performing the experiment, we conducted a pilot test to verify that the tools and overall setup were functioning correctly. The pilot test was sent to the employees at CompanyX who were aware of the experiment, and we asked one employee from CompanyX to click the link to ensure that the interaction was recorded correctly. The phishing campaign was launched on June 13 2022 at 1:06 PM and data collection was interrupted after one day due to the company's reaction to the attack (see Sec. 8.4.1).

8.3.4. Interviews and data analysis

After the field experiment, we conducted interviews to collect participants' experiences, thoughts, and emotions that influenced their response to the phishing attack. We interviewed nine employees of CompanyX. The interviews were conducted over a period of three weeks starting from the week after the phishing campaign had ended. This was to capture the participants' reactions to our campaign as soon as possible.

Interview design

Interviews are a qualitative technique suited to answer exploratory and descriptive questions [230], such as RQ2 in Sec. 8.3. We conducted interviews to gain a more profound understanding of the factors that influence the cognitive processes involved in detecting and responding to tailored phishing attacks. Moreover, the interviews aimed to understand the community defense mechanisms within the company.

To achieve these goals, we developed a semi-structured interview protocol comprising a set of standardized questions that were posed to all participants. Supplementary questions were also asked in the event of any uncertainties or based on the participants' responses. The interviews covered the following topics:

1. The security awareness of the employees.
2. The rational and emotional response upon reading the phishing email.
3. The emotional drive that led employees to report the phishing email.
4. The behavior fostered by the tailored nature of the attack.

The complete list of questions used in the interviews is given in Appendix C.3.

Interviewing participants

The participants were interviewed online for 30 to 45 minutes, and the sessions were recorded for later coding and analysis. Participants were not informed of the specific questions beforehand to minimize bias. The recordings were kept anonymous and stored in a secure facility for the study duration after which they were destroyed. The interviews were transcribed verbatim, and the answers to our predetermined baseline questions, as listed in Appendix C.3, were used in the subsequent coding process.

Open coding

The coding process was performed in multiple sessions, following the card sorting procedure outlined in [390]. This process involved isolating relevant segments from the interviews and printing them for coding. A mixed bottom-up and top-down approach was adopted: the top-down approach was based on the topics relevant to the research, as listed in Section 8.3.4, and the bottom-up approach allowed for the identification of themes from the data. The coding process involved identifying significant segments and assigning appropriate codes, collating codes into potential themes, and applying multiple rounds of review to ensure the coherency and consistency of the themes [50]. The initial coding was carried out by three researchers, with two researchers completing subsequent rounds of review and analysis. The coding process was carried out on printed cards and post-its, with the help of a virtual whiteboard shared among researchers.

The collected data were structured in *topics*, *categories*, and *themes*. A *topic* is a specific code derived directly from the participants' responses to our interview questions; multiple related topics form a category. A *category* is a set of closely related topics that form a sub-theme; such a sub-theme is a first-level grouping that emerges from the bottom-up analysis during the coding of the interviews. A *theme* is a high-level collection of categories that characterizes a particular concept from the top-down approach (i.e., stemming from the structure of the interview questions) and a recurring pattern as it emerges from the bottom-up approach (i.e., stemming from a relationship of categories). Thus, the *topics*, *categories* and *themes* form a thematic 'map' or hierarchy — an overall conceptualization of the data patterns, and relationships between them — where the topics are the atomic elements closest to the data, as opposed to the main overreaching themes farther from it [50].

Results interpretation

To interpret the themes and categories identified using the coding process, we employed a two-step approach. First, we provided an overview of the categories and themes and their interconnections, highlighting the relationships between them to give a semantic description or 'surface' meaning of the data (cf. Sec. 8.4.2). We further interpret our results on the basis of a framework of cognition presented in [57] to reconstruct the underlying thought processes and the influencing factors (cf. Sec. 8.4.2). The framework in [57] specifies a basic structure of human information processing and is meant to analyze the effects of phishing attacks on human cognition. By linking the identified themes to the framework, we can isolate themes and categories of topics that may contribute to the thought process of the participants.

8.3.5. Ethical considerations

This research was executed under ethical approval from our institution's ethical review board under approval number `ERB2020MCS13` and by the management of the company. For the phishing experiment, we followed best practices concerning consent waving and user debriefing [293]. 'Submitted' user passwords as well as the association between user identities and their real names were neither transmitted nor saved by the system. Interviews were

Table 8.1: Phishing experiment results

Response	Count
Emails sent	30
Clicked on the link	5
Filled in random email addresses	2
Requested a working URL the next day	1

carried out according to common guidelines[2] and recordings and transcripts were stored securely and anonymized whenever possible to maintain confidentiality and protect privacy.

8.4. Results

8.4.1. Phishing field experiment

The outcome of our field experiment was surprising. Overall, the company reacted quickly against the phishing attack. In particular, multiple emails reporting the phishing attempt were sent to all employees within ten minutes after our phishing emails were sent. Moreover, the CEO, who was impersonated in the email, was contacted multiple times within minutes after the campaign launch to verify the email's authenticity.

The results of the field experiment are reported in Table 8.1. We can observe that six out of 30 employees who received the phishing email interacted with it. Specifically, five employees clicked on the link in the phishing email. Among these employees, two provided fake login details instead of their company email addresses. Further analysis revealed that this was done to investigate the functionality of the phishing website. Additionally, one employee sent a follow-up email to the email address used in the phishing attack after the campaign had ended, inquiring about the functioning of the link in the email as it was not operational at that time. This might indicate that the employee did not detect the phishing attempt or read the warning messages circulated by their colleagues. Overall, these results should be considered inconclusive due to the warning messages sent out within the company shortly after the campaign's launch.

Two main factors have contributed to this outcome. First, some employees received a warning from Google, shown in Fig. 8.1, which was not triggered during the pilot of the experiment. Second, the targeted company is a research and development firm whose employees are highly-trained and have a technical background. During the subsequent interviews (cf. Section 8.4.2), it became clear that employees have a high level of cybersecurity knowledge, which they have acquired from various sources. This topic is further explored in Section 8.4.2.

Overall, it is not possible to measure the effectiveness of using public data in a tailored phishing email based on our experiment. Nonetheless, the swift response to our phishing campaign is worth investigating to understand the motivations that drove employees to act

[2]Ethical Guidelines for Online Interviews - https://voices.uchicago.edu/202003sosc20224/2020/06/25/ethical-guidelines-for-online-interviews/

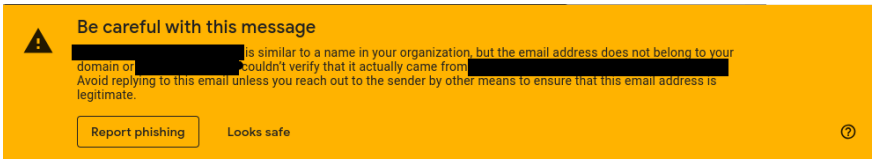


Figure 8.1: Warning displayed by Google to some targets.

Table 8.2: Interviewees and their version of the phishing email

ID	Email version
P_1	Personalized
P_2	Not personalized
P_3	Not personalized
P_4	Personalized
P_5	*The email was not sent (sysadmin)*
P_6	Not personalized
P_7	Not personalized
P_8	Not personalized
P_9	Personalized

quickly to protect their organization from the targeted attack. To this end, we conducted interviews to better understand the reporting process (both to colleagues and the CEO) upon detecting the tailored phishing attack and to shed light on the community defense mechanism in place.

8.4.2. Interviews

In total, nine employees were interviewed. We recruited eight participants from the targeted employees in the phishing field experiment; additionally, we interviewed a system administrator that was excluded from the field experiment. Table 8.2 shows the version of the email received by each interview participant.

Identified themes and categories

In total, we identified 26 categories of topics grouped over six themes, namely *Action*, *Awareness*, *Detection*, *Reaction*, *Reporting (expectations & reasons)*, and *Tailoring effects*. A detailed overview of the identified themes and categories is provided in Table 8.3. The table also reports the number of times each category and theme was mentioned by the participants. The results of the open coding indicate that all six themes were consistently mentioned by the majority of participants, potentially, due to how the interviews were structured. However, variations in responses were observed primarily in the underlying categories.

Awareness. From an analysis of the themes and categories, we find that phishing awareness was primarily obtained by the interviewed employees from previous 'Experience', 'Learning & training', or their combination. In some cases, Awareness conditioned the detection of the

Table 8.3: Identified themes and categories

Theme	Category
Awareness (9)	Source of awareness: Learning & training (5) Source of awareness: Experience (6)
Detection (8)	Technical (3) Expectations mismatch (7) Contextual (3)
Action (7)	Delete email (1) Click (1) Post-detection link investigation (4) Reporting (3)
Reaction (9)	Emotional response: Negative (3) Emotional response: Neutral (2) Emotional response: Positive (4) Mixed response (4) Rational response: Reasons to believe (1) Rational response: Reasons to not believe (3) Rational response: Lack of consequences (1) Rational response: Need for confirmation (4) Rational response: Certain of phishing (5) Rational response: Reasons to action (1) Rational response: Other (3)
Reporting (expectations & reasons) (9)	Reasons for reporting (to email provider) (3) Reasons for not reporting (4) Hypothetical reporting.Report to colleagues (2) Hypothetical reporting.will not report (3) Hypothetical reporting: Other (2) Reporting expectations: Presence of expectation (5) Reporting expectations: Lack of expectation (3)
Tailoring effects (9)	Effect on behavior: No reporting for generic (6) Effect on behavior: More interaction with tailored (2) Effect on behavior: Always reporting (1) Response to being the only target: Emotional response (2) Response to being the only target: Rational response (7) Response to being the only target: Mixed (1)

phishing emails; for example P1 states: *"[...] I did work in the security side of things 20 years ago. [...] I'm an educated person in this regard. […] I do not click on links without checking where it leads, it's standard practice*".

Detection. The Detection theme consists of three categories, namely 'Technical', 'Expectations mismatch', and 'Contextual', highlighting the mechanisms and methods employed by the employees to detect our phishing campaign. 'Technical' clues include characteristics such as the sender's address or domain registration date, (e.g. P1, mentioned above, detects the phishing attempt by inspecting the actual link). 'Contextual' clues relate to contextual factors such as warnings or the language used in the email (e.g., P5, "*Because why would two Dutch people e-mail to each other in English*?"). This may reveal a contrast between the general internal communications in the company (which are run in English), and specific

expectations for communication with the CEO who, according to the participant, was expected to be in Dutch. 'Expectations mismatch' clues represent discrepancies between what is expected and what is observed, such as the tone of the email or receiving personal gifts from the company, (e.g., P3: *[...] we would have, like, a choice between a couple of things and one of them will be, like, very geeky*). Notably, Table 8.3 shows that 'Expectations mismatch' is the category mentioned more frequently as the detection method (7 participants), suggesting this as a common detection method.

Action. A common action taken in response to our phishing email was 'Post-detection link investigation', which includes topics such as investigating the link by filling in fake credentials or hovering over it, while one participant reported having clicked the link because they thought it was a legitimate email, i.e., P8: "*I thought it was legit when reading it on my phone.*". Three respondents indicated that they *reported* the email to the sender, to other colleagues, or even to Google, such as P7: "*reporting it to Gmail makes it more aware that it is a spam e-mail*".

Reaction. Reaction encompasses the various explicit thoughts and emotional responses elicited by receiving the phishing attempt. 'Emotional response' varied from mostly positive, such as "*I always enjoy the gesture [the pretext]*" (P3) and "*This is a nice one [the phishing email]*" (P6), to some negative and mixed as: "*I do wish that I did not click the link [to investigate]*" (P7) and "*Wow, this is the most worthless present ever* [...] *then I thought* [the phishing attempt] *was actually kind of funny*" (P5). Among the explicit or 'Rational responses', some interviewees reported that their initial thoughts were to seek confirmation from others or by other means about the phishing email ('Rational response: Need for confirmation-Reasons to not believe' in Table 8.3). Some were certain of phishing and decided to wait or let someone else handle it: "*I don't trust this, let's wait a couple of days.*" (P2) and "*I figured it would get to* [CEO] *quickly anyway, I wouldn't need to help.*" (P3). Others decided to act by alerting colleagues (e.g., P7: "*I did report. I still left it open 'cause I thought I keep it now like a kind of a fun souvenir.*") or confirm with the intended sender directly: "*[quoting a message to the CEO] it looks like somebody is impersonating you, please check if you are yourself"* (P6). Overall, the interviewees were able to recall detailed thoughts when they were sure it was phishing, but also when they needed confirmation or had reasons to not believe the email, with a total of 12 mentions for these categories (cf. Table 8.3). On the other hand, other reactions, such as emotional reactions, were more sparse and were reported vaguely with short descriptions such as ''*it felt weird*", "*this is funny, just let it pass and let's wait*", "*I'm not sure if I clicked*" or even "*excited we are phished*". This may be an indication that participants who detected or were suspicious about the phishing email engaged in a more cognitive effort to process the situation as they recalled more details when asked the relevant questions.

Reporting (expectations & reasons). Theme Reporting (expectations & reasons) groups together participants' reasons and expectations to (not) report the phishing attempt, as well as reporting in the future ('Hypothetical reporting' in Table 8.3). Among the reasons to report, interviewees cited to protect others, such as P6, "*I want to prevent people from clicking*", and P1: "*I'm assuming that if a number of people say that's phishing, that at least everybody gets the alert*". Other participants' reasons to not report indicated they did not know how to do it, thought that someone else should do it (P3, "*I pretty much expected that multiple people notified* [it])" or simply preferred to do nothing, for example, because "[reporting] *hasn't any*

effect [in Google]" (P2). When asked about hypothetical future reporting, most of the interviewees remained consistent with their previous answers on reasons to report, such as P1: "*I would [report], I think that more people would do it in the group.*" and P6: "*I think that this type of just sending a quick warning around is easy enough.*" On the same line, participants that responded they did not report were consistent with not reporting in the future, e.g. P2: "*We work with Google and we don't care so* [...] *I wouldn't warn any other colleagues, no.*" and P3: "*I might not have warned my colleagues, probably only* [the CEO]". Similarly, the 'Reporting expectations' from other colleagues, or the lack thereof, were coherent with previous answers, i.e., those who said they would (not) report, do (not) expect others to report.

Tailoring effects. Within theme Tailoring effects – the effects of the tailored nature of the attack – two categories were mentioned more frequently than others. Specifically, participants stated that they were less likely to report generic phishing emails, for example, P6: "*if it was that generic, then I would just have ignored it*" and P7: "*I would be less likely to inform my colleagues that we are being scammed*". When asked about their feelings about being the only target of the campaign, they primarily provided rational responses, pointing at the unexpectedness of such a possibility: "*I would be surprised because I don't have any social media.*" (P5) or "*I would be very confused as to why [they] went through all those lengths to only target me*" (P3). This observation suggests that participants would be more likely to engage in rational decision-making when faced with targeted phishing attempts (perhaps, due to the interest induced by such a 'rare occurrence'), and that the perceived level of personal relevance of the phishing email may influence their likelihood to act on it, for example, by reporting it: "*if that type of data was in that e-mail, then not only I would have been triggered and warned everybody, but then I would have gone into a full counter-attack mode*" (P6).

Results Interpretation

To provide a more in-depth analysis of the underlying factors that may condition interviewees' motivations and decisions, we consider the cognitive processing for phishing reported in [57] (cf. Chapter 3) and structure our analysis over three dimensions: 1) *detection*: a participant reads the email and may or may not detect it as phishing; 2) *reaction*: various reactions can follow, such as a positive emotional reaction or more rational thought; 3) *action*: an action can take place, such as deleting and/or reporting the phishing email. We follow this structure to reconstruct the connection between these three steps and the identified themes. We consider a connection between a theme and a step to exist when a theme's topics are directly related to a step (e.g., a detection occurs due to a mismatched URL) or when a theme's topics describe or characterize a step (e.g., previous experience allowed to spot a mismatched URL).

Through this process, all six themes appear to be related to the basic cognitive process (**Detection**, **Reaction**, **Action**, **Awareness**, **Tailoring effects, and Reporting (expectations & reasons)**) and were schematized in Fig. 8.2. It is worth noting that the schema in Fig. 8.2 serves as a simplification and abstraction of the complex interactions and factors that were identified in the data.

Themes **Detection**, **Reaction**, and **Action** naturally fit the basic cognitive process steps of Fig. 8.2 as the participants' mentions (belonging to those themes) describe how they detected,

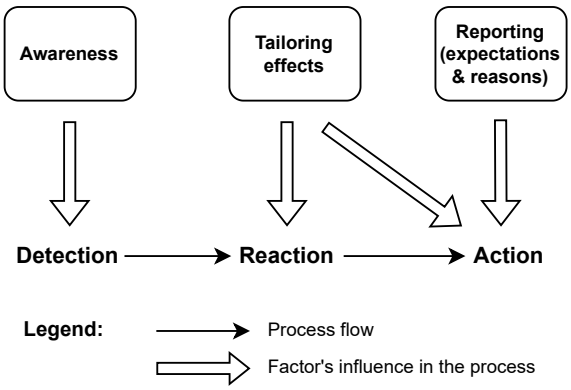


Figure 8.2: Illustration of participants' cognitive steps and potential contributing factors.

reacted and acted upon receiving the email. For example, the category 'Rational response: Need for confirmation' (under **Reaction** in Table 8.3) contains users' reaction of seeking a confirmation that this was indeed a phishing email, P1: "*if somebody else would have been in the office, I would have asked it*", thus the **Reaction** theme corresponds to the Reaction step in Fig. 8.2.

Themes **Awareness**, **Tailoring effects**, and **Reporting (expectations & reasons)** contain participants' mentions of additional aspects that affected or characterized their detection, reaction, and action steps of the process. Therefore, they can be described as contributing factors to the overall cognitive process experienced by the targets during the phishing campaign. These themes do not represent clear steps that individuals actively went through, but rather provide insight into factors that may have influenced their behavior and decision-making.

The contributing factor **Awareness** encompasses the 'Sources of awareness' among employees, such as the previous experience of P6: "*The various ways of trying to attack people through cyber means are familiar to me, so I'm in that sense an educated person in this regard.*" (P6). Thus, **Awareness** conditioned how P6 detected the phishing attempt (*Awareness -> Detection* in Fig. 8.2): "*I do not click on links without checking where it leads, it's standard practice.*" (P6)

Theme **Tailoring effects** illustrates how the tailored nature of the attack impacted employee actions by encouraging them to investigate or report the email (*Tailoring effects -> Action* in Fig. 8.2). For instance, answering to Q4a, P4 stated "*I would probably be more inclined to just share it* [the email] *with the people I work with*" and P7 stated "*I would have instantly sent it to spam and then reported it*". Similarly, **Tailoring effects** affected the Reaction to being the only target (*Tailoring effects -> Reaction* in Fig. 8.2). For instance, answering to Q4b, P3 mentioned that "*It will be weird to only target me*", P5 that "*I would be surprised because I don't have any social media. So, the only way you could know I work here is by how long my picture has been on the website, so you probably have used Wayback Machine and looked at snapshot*", and P8 that "*It's a weird feeling, not sure what the other party wants of you, do they want money, do they want information about your thesis subject?*".

Finally, the contributing factor **Reporting (expectations & reasons)** appears to affect the motivations for reporting (or not reporting) the phishing email, hypothetical reporting in a similar future scenario, and reporting expectations among colleagues in the future (*Reporting (expectations & reasons) -> Action* in Fig. 8.2). For instance, P5 stated that "*I expected someone of our own to click the link and fill their credentials. And my initial reaction was, we should probably send a take-down notice.*" and P8 that "*I would also expect people to sort of send that e-mail [the report], I might now also do it myself*".

Overall, the conceptualization in Fig. 8.2 has the goal to illustrate that factors, such as target-relevant information and social dynamics, might have considerate effects on phishing response in (small) organizations. The positive role of awareness in detecting (generic) phishing has been the focus of a large proportion of research [147], however, whether user awareness is as effective in more advanced scenarios (e.g., [323]) still remains an important question to address. Another hypothesis stemming from our results is that the tailored nature of attack artefacts can influence an individual's reaction to it (spark interest, e.g. [120]) and, eventually, lead them to act (investigate, report). Finally, among factors that may push individuals to act, motivations to report (similarly to Chapter 7) and expectations from their colleagues appear important drivers for reporting as well [195]. Investigating such hypotheses in future studies might shed light on the effects of tailored attacks, test the findings of this and related work and phishing mitigation strategies.

8.5. Discussion

Overall, the detection and swift reaction to our simulated attack might be attributed to several factors: attack-related factors, e.g., the mismatch of attacker assumptions and target characteristics; contextual factors, e.g., the warning message displayed in Gmail (see Fig. 8.1); and target-related factors, such as employees' awareness and social dynamics of a small tech company. A lesson drawn from our experiment is that it may be very difficult to execute tailored phishing attacks in an environment where members know and help each other, in contrast with 'siloed' structures typical of larger enterprises.

Detection

Extant research has shown that employees are typically able to spot inconsistent patterns to detect tailored phishing attacks due to their experience in the context of the organization [54]. Whereas employees at large organizations may need a long experience in the context of the organization to be able to spot tailored attacks (cf. Chapter 5), members of the small enterprise under study appear to do so regardless of their tenure, perhaps, due to the overall limited size of the company where 'everyone knows everyone'. In our case, the participants used their acquaintance with the company's CEO to determine whether the email was unusual based on factors such as structure, content, and language. For instance, the email signature was reported as an example of misalignment with the participants' expectations: P4 said *"I mean the signature is not the signature we typically use. It was an older version, I think."*. The expectations mismatch mechanism can be seen in the context of a cognitive modeling of the phishing attack (cf. Chapter 3): the mismatch between the attacker assumptions on the targets and the real target characteristics (e.g., salutation, language, signature)

can be more easily achieved in small and connected contexts (such as a small enterprise). As an implication, future experiments with tailored attacks may need to assure that their modeling of a realistic attacker can match the attacker's assumptions and target parameters, while organizations may want to maximize the mismatch effect when providing training and awareness campaigns to their employees.

Response

The tailored nature of the attack sparked curiosity, concern, and even amusement among participants, leading four of them to investigate the contained link. However, all of them reported having experience with phishing and still clicked the link to investigate. Interestingly, the participants who sent the warning about the phishing campaign said that the email triggered their interest because the phishing email was a 'quite good one'. This surprisingly swift defense reaction of the employees hampered our efforts to investigate the effects of the tailored phishing attack in the first minutes of delivery. This appears to be very dissimilar to and potentially much more effective than what happens in large organizations: as seen in Chapter 7, the group defense mechanism of several employees sending a warning to their colleagues was faster than common response and containment procedures in organizations (often equipped with staffed IT departments) [195, 352]. This may be attributed to the network effects of a small organization mitigating bystander effects. Indeed, six out of nine participants reported doing nothing and relying on someone else to handle the situation (e.g., P4: "*We have like the two cyber security guys that are on top of this, they will trace this guy*"), but the other three alerted the whole company immediately (one did so inside the Gmail interface). Existing literature shows that the detection of a phishing attempt does not always lead to its reporting, as users may have different mental models and approaches to decision-making when it comes to reporting or not reporting phishing emails (see Chapter 7) [195]. Some users report an email, even if they are unsure whether it is an actual phishing attempt; others do not report phishing emails unless they perceive them as particularly dangerous or sophisticated. Others may only report phishing emails if they believe the security of their colleagues is at risk, or might not because they expect someone else to do so. These findings imply that there may be a need for security awareness programs to emphasize the underlying reasons why reporting is important, for example, by presenting real case studies that employees can relate to [204]. Future research may investigate how larger organizations can achieve similar effects, in terms of 'group defense', against sophisticated phishing attacks as in our case study.

8.5.1. Limitations and threats to validity

Construct validity. The presence of confounding variables may have influenced the results, in particular, the phishing campaign was disrupted by the issuance of two warnings shortly after the campaign launch, which affected the experiment's results.

External validity. The sample size in our experiment is relatively small, consisting of 30 employees for the field experiment and 9 for the interviews. Moreover, we performed the study in a research and development SME which employees are highly-trained and have a technical background. Therefore, our findings might not be applicable to other types of organizations.

Further experiments, with larger sample sizes and targeting SMEs in different sectors, should be conducted to confirm our findings.

Researcher and respondent bias. The threat of researcher bias was mitigated by carrying out the open coding of the interview transcripts by two authors with an additional review iteration with all authors. As the interview questions were open-ended and intrinsically not sensitive nor controversial, we believe that respondent bias is not a significant threat to validity.

8.6. Conclusion

In this work, we conducted a field experiment and subsequent interviews to study the effects of tailored phishing campaigns targeting SMEs. Our campaign was detected shortly after its launch, as all employees were promptly alerted. Driven by this unusual response, we investigated the cognitive processes concerning the detection and response to tailored phishing attacks. Our findings show that expectation mismatch is a key factor in detecting advanced phishing emails. Moreover, the tailored nature of the attack provoked some participants to quickly alert the rest of the group, thus mitigating the bystander effect of other employees, which is often observed at larger organizations. The swift defense reaction may be attributed to the network effects of a more connected community, as opposed to 'siloed' structures typical of larger enterprises.

Our investigation into individuals' reactions and perceptions to the phishing campaign contributes to answer RQIV. Despite the small sample size in our study, these findings provide potential hypotheses for future research and potential indications for the design of more effective awareness training. In Chapter 9, we carry out a user study to evaluate the relationships of human traits, attitudes and beliefs with intention to report phishing at organizations.

9

Traits, beliefs and attitudes that influence phishing reporting

Prior chapters identified phishing reporting as a key, largely untapped resource to mitigate phishing threats at organizations. However, phishing reporting practice suffers from very low reporting rates and users are often unaware of the value of reporting. Whereas it is known that phishing reporting behavior is affected by a number of 'human factors', such as personality traits, attitudes towards the organization and colleagues, the overall picture remains fragmented, and therefore not yet actionable (see RQIV). To provide a cohesive view of individual and organizational factors affecting individuals' intention to report, we build a theoretical picture of their effects and constructs and develop, model, and empirically evaluate (by means of an online questionnaire with 284 respondents) the resulting hypothesis structure. We discuss both theoretical implications of our findings and research directions for practice at a research and organizational level.

This chapter is originally published as I. A. Marin, P. Burda, N. Zannone, and L. Allodi, "The Influence of Human Factors on the Intention to Report Phishing Emails", In *Conference on Human Factors in Computing Systems* (CHI), ACM, 2023, pp. 1–18

9.1. Introduction

Part of the reason why phishing attacks remain so successful is that current phishing countermeasures mainly focus on detection techniques based on spam filters and blacklisting of phishing domains, which have proven insufficient particularly to detect more targeted variants of these attacks [15]: spam filters are oftentimes incapable of detecting well-engineered, credible phishing attacks, and the speed by which these achieve their objectives makes blacklisting simply too slow to be effective in time (i.e., before the attack targets have been victimized).

Chapter 7 has stressed the importance of phishing reporting as an additional [22], potentially fast, crowd-source based countermeasure to timely react and mitigate phishing attacks. Phishing reporting is a practice adopted in most organizations (either by internal means or by reporting tools provided by major software platforms, such as Microsoft's Office365) that counts on the organization's employees to report suspicious messages, generally in the form of emails [22], to the organization. Reporting is increasingly more prevalent in phishing awareness material and training exercises, yet phishing reporting rates remain steadily below 10% [195, 352].

How to maximize and fully exploit the additional line of defense represented by phishing reporting remains an open question. Human factors such as personality traits [100], employees' attitudes towards the organization [261] and towards their own colleagues have been shown to play an effect on individuals' cyber security behaviors, in general, [166], but the overall picture remains fragmented, and therefore not yet actionable, in the literature [195]. In this chapter, we argue that a full picture can only be derived by looking at both an individual (e.g., personality traits) and organization (e.g., relating to employees' security assurance behaviors and compliance to security policies) perspective simultaneously and by focusing on phishing reporting behaviors (as opposed to generic cyber security behaviors). Moreover, the lack of a single theoretical picture linking together different relevant theories in a cohesive framework limits the reusability and actionability of findings. Importantly, 'extra-role' behaviors (e.g., those not necessarily mandated or motivated by an organization's policy) have not yet been considered in the picture.

To address these gaps, in this study we unify different perspectives pertaining to human traits and organizational cyber security behavior towards both the organization itself and individuals, and evaluate their joint effect on the intention to report phishing emails. Stemming from RQIV, we formulate the following research question:

RQ: *How do human factors, pertaining to the individual and the organizational levels, influence the intention of reporting suspicious phishing attacks?*

To answer our research question, we first evaluate the extant literature to identify theories and factors pertaining to individual and organization-level cyber security behaviors, and personal characteristics such as beliefs and the 'Big Five' personality traits. Based on these, we derive our hypotheses and construct a unified theoretical model of the human factors affecting an individual's intention to report phishing emails. The model and hypotheses structure are used to create an online questionnaire aimed at empirically evaluating and quantifying

the link between the identified factors, their relation, and their joint effect on individuals' cyber security behavior and their intention to report phishing emails. We conduct the survey on Amazon Mechanical Turk (AMT) with 284 participants. The main contributions of this chapter are as follows:

- The developed hypothesis structure and related theoretical model address the fragmentation of the extant literature by providing a cohesive picture of individual and organizational factors, affecting individuals' cyber security behaviors and their intention to report phishing emails, and their interplay.
- Our empirical evaluation shows that accounting for different types of human factors (personality traits, beliefs, attitudes towards the organization and co-workers) at both individual and organization levels provides a more comprehensive understanding of their effects on individuals' positive cyber security behaviors and intention to report phishing emails. For example, emotional stability and extraversion traits are not aligned with previous results on cyber security behaviors, potentially due to the inclusion of other factors such as organization-related factors.
- Our evaluation also shows that the human factors that influence an individuals' cyber security behaviors, in general, might differ from the factors influencing a specific cyber security behavior such as the reporting of phishing emails. For instance, conscientiousness and extraversion appear to have a strong relationship with generic cyber security behaviors, but this does not translate to the specific behavior of reporting; to the contrary, we observed that altruism only influences reporting.
- The understanding of the effects of human factors has implications at both a theoretical and practical level and can help organizations to improve their overall security posture. Our findings can support researchers and practitioners in the design of better training practices and awareness programs and can help organizations to create a security culture. For instance, our results show that high-sportsmanship individuals, who usually tend to avoid filing complaints, are characterized by a lower intention to report phishing emails; a training program may mitigate this effect by stressing the relevance of phishing reporting in terms of increased overall security.

The remainder of the chapter is structured as follows. The next section introduces the relevant theoretical background on information security behaviors and their relation with human factors. Section 9.3 presents our hypotheses and Sections 9.4 presents the methodology to test those hypotheses. Section 9.5 presents the results and Section 9.6 discusses the implications and relevance of our findings at both a theoretical and practical level, as well as the limitations of our study.

9.2. Background and Related Work

To protect their sensitive information and assets, organizations not only employ various types of security mechanisms but also take measures to improve their security posture. To this end, organizations often engage their employees with security *training* programs to educate them, for instance, on how to detect and report phishing scams, and with phishing *awareness* programs to ensure that they become familiar with how phishing attacks are deployed, recognize when they are the target of a suspicious phishing email and react accord-

ingly [285]. More in general, organizations often aim to create a security *culture* providing their employees a pattern of shared basic assumptions and principles that work well enough to be considered valid and, therefore, shape employees' perception and behaviors adopted within the organization [308]. Therefore, organization culture, training and awareness are clearly involved in shaping employees' positive cyber security behaviors and may be valuable tools to mold phishing reporting behaviors. However, their effectiveness and uptake often depends on the employees' experiences, personality traits, characteristics and beliefs.

The cognitive science field, applied to the InfoSec domain, provides the foundation of the current study as it identifies the human factors involved in the cognitive processes that emerge when encountering a phishing email. A number of theories have been proposed to understand the relationship between human factors and the intention of an individual to engage in performing an action [8, 298]. In this work, we apply these theories to positive cyber security behaviors with a particular focus on users' intention to report phishing emails. Next, we provide an overview of the most influential concepts that aim to explain InfoSec-related behaviors, that will form the basis of our research model.

Table 9.1: OCB Characteristics

Class	Characteristic	Description
OCBO	Civic Virtue	This trait represents an individual's non-obligatory concerns regarding the welfare of the organization as a collective, creating a sense of community by including their voluntary active participation towards solving existing issues and improving the organization's processes.
	Leader Support	This characteristic covers the employee's perception of positive behaviors received from their superiors. Such behaviors include both task-oriented actions (e.g., receiving help with ongoing projects, appropriately setting goals and deadlines), and socio-emotional actions (e.g., effective communication and interaction) [23].
	Organizational Commitment	This characteristic refers to the *Affective* Organizational Commitment, which captures an individual's connection with the organization's values and objectives, while displaying similar views that lead to beneficial behaviors.
	Sportsmanship	This characteristic covers an individual's tolerance for less-than-ideal situations encountered within the organization, when, even though they might not fully agree with or be aware of the circumstances, employees do not display complaining or negative behaviors.
	Conscientiousness	This characteristic represents a narrower form of generalized compliance [203], involves self-discipline, and refers to employees whose helping behaviors exceed simply adhering to the rules and obligations prior established by the organization.
	Job Satisfaction	This characteristic indicates the employee's contentedness with their workplace, including their perception of different aspects related to the type of job they perform, or the organization they are associated with.
OCBI	Altruism	This characteristic defines the extra-role behaviors that are directly intended to assist or provide support to others in an organizational setting (e.g., helping co-workers by taking some of their tasks, or volunteering to perform an action that is not required) [203].
	Courtesy	This characteristic refers to performing considerate actions with the goal of avoiding problems that could occur or impact the organization (e.g., "Is mindful of how his or her behavior affects other people's jobs") [186].

9.2.1. Reasoning on InfoSec-related Behaviors

Cyber Security Behavior Classification

Guo [135] proposes a framework to classify employees' InfoSec-related behaviors observed in organizations. These behaviors are structured into four categories: *Security Assurance Behavior (SAB)*, *Security Compliant Behavior (SCB)*, *Security Risk-taking Behavior (SRB)*, and *Security Damaging Behavior (SDB)*. The first two categories, SAB and SCB, focus on the desired behaviors that an organization should encourage, while the latter two categories, SRB and SDB, are the behaviors that an organization should prevent. Specifically, SAB describes behaviors that an employee actively performs with the intention to protect the organization's systems (e.g., reporting security incidents), whereas SCB describes both intentional and unintentional behaviors aimed to comply with an organization's Information Security Policy (ISP). On the other hand, SRB describes intentional behaviors that could harm an organization's data security (e.g., writing sensitive data on paper), while SDB describes intentional behaviors of an employee that directly damage the organization (e.g., data theft). As here we focus on phishing reporting, in this work we only consider positive behaviors, namely SAB and SCB, to which we refer as *Positive Cyber Security Behaviors*.

Organizational Citizenship Behaviors (OCB)

OCB refers to "individual behavior that is discretionary, not directly or explicitly recognized by the formal reward system, and that in the aggregate promotes the effective functioning of the organization" [260, p. 86]. Since its introduction, OCB has attracted much attention from the research community to identify and analyze the relationship between various behavioral dimensions, known as predictors, that impact OCB. Following the division proposed in [373], OCB is refined into two distinct subgroups of characteristics, depending on the target of the behavior: OCB directed towards the Organization (OCBO) and OCB directed towards Individuals (OCBI). In the InfoSec context, OCBs influence both positive and negative cyber security behaviors [100]. In particular, OCB is linked to behaviors supporting and reducing potential InfoSec harm in an organization (SAB and SCB).

9.2.2. Human Factors

Next we discuss, from the extant literature, the most relevant human factors affecting the intentions and actions of individuals from a cyber security perspective.

OCB Characteristics

Several studies have investigated OCB characteristics and their implications on behaviors related to InfoSec. An overview of these characteristics is presented in Table 9.1. Helping/-supporting behaviors towards the organization and co-workers are generally beneficial to the organization's cybersecurity posture. Within OCB characteristics directed towards an organization (OCBO), *Sportsmanship* is known to have a notable impact on the overall predisposition of individuals to be helpful in organizational contexts [203], where individuals with a high level of Sportsmanship are less prone to have negative reactions and complain

Table 9.2: The Big Five dimensions of personality

Dimension	Description
Agreeableness	This dimension includes interpersonal characteristics related to being courteous, flexible, trusting, cooperative, tolerant, and forgiving [41].
Conscientiousness	This dimension involves dependability, namely being careful, thorough, responsible, organized, and planful [41].
Openness to Experience	This dimension reflects an individual's imaginative, cultured, curious, original, broad-minded, intelligent, and artistically sensitive aspects [41].
Extraversion	This dimension includes interpersonal characteristics related to sociability, gregariousness, assertiveness, talkativeness, and activeness [41].
Emotional Stability	This dimension covers an individual's emotional adjustments, observing the lack of anxiety, anger, embarrassment, worry, insecurity, and impulsiveness [41].

about current issues as they are more oriented towards future improvements and their contribution to these changes [299]. Employees' purposeful help and support directed at the organization when solving encountered issues is influenced by their level of *Civic Virtue*, which influences the direct contribution of employees in the protection of the organization they work for [297]. Differently, the motivation behind employees' beneficial behaviors is driven by *Organizational Commitment*, which is typical of employees that share the organization's views and ideals (*cf.* culture, above) [285].

At the *individual* level of OCB, *Altruism* and *Courtesy* are OCBI characteristics that relate to actions directly intended to help co-workers [162, 329] and contribute to the smooth functioning of the organization [329]. In particular, it has been shown that Altruism improves employees' overall performance in executing their daily tasks and the collective efficiency of the organization [162]. Differently, employees showing a high level of Courtesy are inclined to perform actions that help avoid or mitigate potential issues, cautiously engaging in any behaviors that may harm their co-workers [299].

Security assurance and compliance behaviors include actions that individuals perform with the aim of protecting the organization from potential security attacks. These behaviors are often associated with a high level of Conscientiousness: conscientious individuals tend to go beyond the minimum requirements and actively engage in security behaviors [141, 162, 299]. This characteristic influences an individual's work ethic and behavior consistency; related to email usage, a high level of Conscientiousness leads to the inclination to regularly check emails and thoroughly evaluate the information of the received emails, resulting in a lower susceptibility to phishing attacks [195]. However, when this repeated behavior is exhibited by individuals with low emotional stability, it may lead to strong email habits [356] such as constantly monitoring email and regularly engaging with links in emails they receive.

The Big Five Theory Characteristics

The Big Five Theory aims to measure and interpret individuals' personality variations, guided by five defined factors characterizing independent human cognition dimensions [227, 253]. The five dimensions of personality traits are: *Agreeableness/Likability*, *Conscientiousness*, *Openness to Experience*, *Extraversion/Surgency*, and *Emotional Stability* (inverse of Neuroticism or Cognitive Impulsivity). An overview of these dimensions is provided in Table 9.2.

Considerable research has been conducted to study the relationship between the "Big Five" personality traits and how these affect the underlying cognitive processes involved in the InfoSec context, including the actions an individual takes when receiving a phishing email [184, 238, 356]. In particular, it has been shown that these dimensions are strong predictors of proactive behaviors related to security assurance behaviors within organizations [100]. For instance, previous research investigated the influence of high cognitive impulsivity on email management, finding supportive evidence that individuals with low *Emotional Stability* tend to be more volatile and, therefore, more likely to engage in behaviors that damage the organization [100], while less impulsive employees are better at evaluating and managing phishing emails [271]. Similarly, individuals characterized by a high level of *Extraversion* are typically able to handle phishing emails and take appropriate actions, including reporting the suspicious email received, even if they are not completely certain it is phishing [271]. Individuals with higher levels of *Agreeableness* and *Openness to Experience* tend to be more receptive of the organization's security training [100]. *Conscientious* individuals are also aware of their organization's rules and regulations and may aspire to adhere to them [100]. Therefore, they usually comply with information security policies and guidelines established by their organization. On the other hand, individuals with low levels of Conscientiousness and Agreeableness may not consider the implications of their behaviors or even deliberately act against their organization if it benefits them [100].

Beliefs

Beliefs refer to subjective interpretations, focusing on one's attitude or perception about a 'truth' that has not been verified, and are known to influence human behaviors. In the context of Social Engineering (SE) attacks within organizational settings, the leading beliefs that affect an individual's behavioral attitude are *Self-efficacy*, *Subjective Norms* and *Habits*, briefly described in Table 9.3.

Self-efficacy and Subjective Norms are beliefs that can influence employees' compliance behaviors within the organization they are part of. Self-efficacy is a dimension of coping appraisals within the Protection Motivation Theory (PMT) [298] and represents "the most powerful predictor of intention to comply with a behavior" [319, p. 218]. This type of belief influences both employees' competence and effort put into the work-related activities, as well as how behavioral patterns evolve [8]. At an organizational level, more experienced and confident individuals are shown to be less inclined to comply with the demands of a deceptive email [380]. In contrast, it was observed that individuals who self-rated their technical knowledge as low are more likely to be subject to phishing [141, 314]. On the same line, the Theory of Planned Behavior (TPB) [8] shows that Subjective Norms, together with an individual's attitude towards a behavior and perceived controls, can be used to predict an in-

Table 9.3: Dimensions of Beliefs

Dimension	Description
Habits	Following the Habit Theory, this type of belief refers to "...learned acts that become automatic responses to situations, which can be functional in obtaining certain goals or end-states" [353, p. 112]. Behaviors performed with repetition under certain cues have the tendency of becoming habitual, where fewer conscious decisions are required.
Subjective Norms	This characteristic concerns the beliefs an individual has regarding a perceived social pressure, namely whether others would approve or disapprove of them performing a behavior. From an organizational perspective, subjective norms are cues that individuals within an organization urge employees to take in order to perform certain actions [8].
Self-efficacy	Self-efficacy represents an individual's confidence that they are capable of performing response behaviors to encountered data security incidents [361]. This behavior can be achieved by adhering to the established ISPs (as a part of SCB), as well as consciously performing actions that lead to the active protection of organizational data (as a part of SAB). More specifically, self-efficacy points to an individual's belief that they can perform anti-phishing behaviors, such as reporting an email received that they have identified as being suspicious.

dividual's intention to perform that behavior with a high accuracy degree in organizational contexts. The relevance of Subjective Norms with respect to an employee's cyber security behaviors has also been studied by Jalali et al. [166], who observed that Subjective Norms positively affect an individual's intention to comply with the ISP of the organization.

Self-efficacy, together with Habits, is also an influential factor for security assurance behaviors. Kwak et al. [195] studied the role of Self-efficacy as a predictor of the likelihood of *reporting* spear phishing emails, showing that individuals with higher levels of Self-efficacy put great effort into reporting, whereas individuals with lower levels of Self-efficacy have self-doubt and do not take further actions. On the other hand, recent research shows that *Email Habits*, i.e. non-intentional automatic behaviors related to email usage, is a predictor of the intention to click on phishing emails [311].

9.2.3. Discussion on related work

Several works in the multidisciplinary domains of InfoSec and cognitive sciences investigate the human factors impacting behaviors aimed at protecting sensitive information or relating to deception. However, previous studies have mainly focused on general positive and negative cyber security behaviors in organizational contexts, covering human factors involved in the beneficial and harmful behaviors individuals perform [100], but little attention

has been given to human factors impacting intention to report. Recent work on reporting has investigated, for instance, whether using the employees as a collective phishing detection mechanism is practical in large organizations [56, 196], the relationship between the believability of a phishing email and the associated reporting rate [179] or how security gamification can improve phishing reporting [171]. A few studies have focused on individuals' behaviors towards phishing emails, for instance, by investigating the reasoning behind why individuals open suspicious emails [166, 311] or the motivations of why phishing reporting is scarce [195]. However, an understanding of which human factors influence an individual's propensity to report suspicious emails is still lacking. Moreover, the vast number of different theories and classifications may hinder the development of a clear overview of the impact that human factors have on phishing reporting. Previous research has mainly focused on the factors that influence in-role behaviors of employees by adhering to ISP [319], while the extra-role behaviors influencing an individual's cyber security actions at their workplace have not yet been assessed.

In this work, we investigate the human factors that influence individuals' cyber security behaviors in organizational contexts and their intention to report phishing emails. To this end, we employ the human factors identified in previous studies, especially concerning actions related to phishing emails, as a baseline for the current research (cf. Section 9.3). Thus, this work adds to the current literature by focusing on the human factors that influence behaviors related to phishing reporting, at *both* organization and individual level.

9.3. Hypothesis Development

In Section 9.2 we identified the human factors and theories that the extant literature has related to cyber security behaviors. We employ those perspectives to derive a theoretical model of human factors, setting the hypotheses tested in this work. To keep the size of the experiment manageable, for each variable category (OCBO, OCBI, personality traits, and beliefs) we select two variables for which the literature reports evidence of their relevance for the security constructs in our model. A more extensive justification of the inclusion of each variable is given in Sections 9.3.1 to 9.3.3.

Figure 9.1 provides a graphical representation of the selected constructs and related hypotheses. The model comprises eight constructs representing the human factors that can potentially influence an individual's cyber security behaviors and their intention towards the reporting of phishing emails. The hypotheses are divided in three groups: the first group investigates which human factors positively affect positive cyber security behaviors; the second group investigates whether positive cyber security behaviors positively influence an individual's intention to report phishing emails; and the third group investigates which human factors positively relate to the intention to report phishing emails.

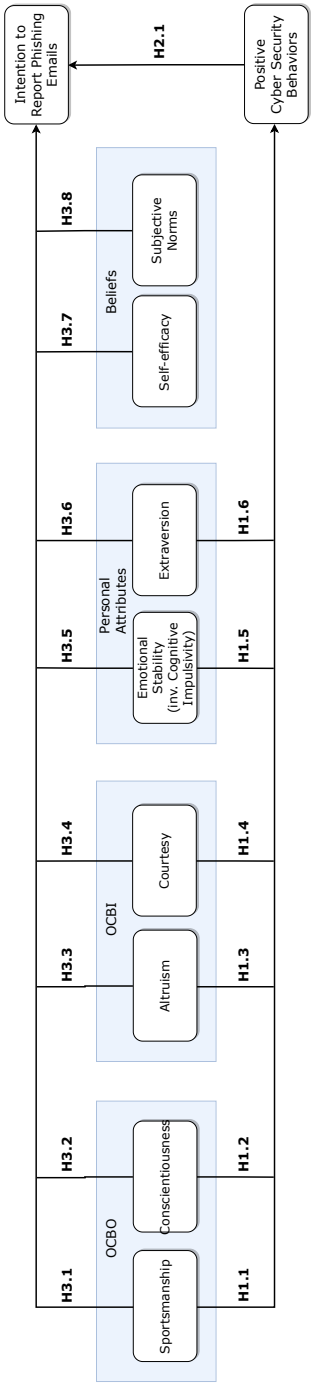


Figure 9.1: Research model

9.3.1. Human Factors affecting Positive Cyber Security Behaviors

Previous work has shown that OCB characteristics and personality traits influence an individual's cyber security behaviors [184, 238], but they were typically studied separately. In this work, we are interested in assessing their combined effects on an individual's positive cyber security behaviors. Among the OCBO characteristics, it has been observed that *Conscientiousness* and *Sportsmanship* are strong predictors of organizational and security behaviors, where individuals with a high level of Sportsmanship are significantly more likely to engage in behaviors that contribute to the good welfare of the organization [299], and Conscientiousness is positively associated with secure behaviours [141] and positively affects how targets respond to received phishing emails [195]. As seen in Section 9.2.2, *Altruism* and *Courtesy* display a strong positive correlation with OCB and are prominent predictors for relevant human behaviors associated with helping coworkers, mitigating and avoiding issues [162, 299, 329] which reasonably fall within the scope of positive cyber security behaviors. In terms of email usage, we aim to test whether individuals who tend to be less impulsive in their decision-making processes and more extraverted, are also more likely to perform proactive security actions as these personality traits are more relevant for an individual's cyber security behaviors compared to other traits. Indeed, as shown in previous studies, individuals with low cognitive impulsivity are more likely to take measures against phishing attacks [63]; similarly, extraverted individuals typically take the appropriate action to handle both genuine and phishing emails [238]. Therefore, we consider *Emotional Stability* and *Extraversion* in our research model. As a result, the following hypotheses are used to test the relationship between the identified factors and positive cyber security behaviors:

H1.1: *Sportsmanship is positively related to an individual's Positive Cyber Security Behaviors.*
H1.2: *Conscientiousness is positively related to an individual's Positive Cyber Security Behaviors.*
H1.3: *Altruism is positively related to an individual's Positive Cyber Security Behaviors.*
H1.4: *Courtesy is positively related to an individual's Positive Cyber Security Behaviors.*
H1.5: *Emotional Stability is positively related to an individual's Positive Cyber Security Behaviors.*
H1.6: *Extraversion is positively related to an individual's Positive Cyber Security Behaviors.*

9.3.2. Positive Cyber Security Behaviors affecting the Intention to Report Phishing Emails

A large body of research has studied individuals' cyber security behaviors within an organization setting, also in the context of phishing attacks. Previous work has often investigated either abstract cyber security behavior constructs, such as PCSBs, or specific cyber security behaviors, such as managing emails or clicking on links. However, in the former case the resulting findings and considerations were often extended and generalised to specific cyber security behaviors, and vice versa for the latter case. Such generalizations and extensions of findings from generic to specific behaviors (or from specific to generic) might not always hold. Moreover, previous studies typically focus on phishing victimization (opening a phishing email or clicking on the link) while an understanding of the actions an individual may perform to *protect* the organization from phishing attacks is far less studied. To fill these

gaps, we investigate whether positive cyber security behaviors influence an individual's intention to report phishing emails, when both extra-role (SAB) and in-role (SCB) behaviors are considered:

H2.1: *Positive Cyber Security Behaviors are positively related to an individual's intention to report emails that they consider to be phishing.*

9.3.3. Human Factors affecting the Intention to Report Phishing Emails

Although several works have studied whether human factors influence an individual's cyber security behaviors (cf. Section 9.2.2), it is still unclear whether they also affect the intention to report phishing emails, especially when reporting requires additional efforts. To this end, in addition to the OCB characteristics and personality traits discussed in Section 9.3.1, we also study how beliefs influence the intention to report suspicious emails. In particular, we study the relation between beliefs and an employee's compliance intentions (in-role behaviors) and analyze whether this knowledge can be extended to the employee's active commitment to protect the organization outside of the described policies and regulations (extra-role behaviors). Among the types of beliefs discussed in Section 9.2.2, we consider *Subjective Norms* and *Self-efficacy* as these human attitude and perception factors can alter an individual's behavioral intentions in the context of reporting suspicious emails. On the other hand, we do not consider *Habits* as they represent behaviors performed with repetition, where fewer conscious decisions are required [356]. The observations above are captured by the following hypotheses:

H3.1: *Sportsmanship is positively related to an individual's intention to report emails that they consider to be phishing.*
H3.2: *Conscientiousness is positively related to an individual's intention to report emails that they consider to be phishing.*
H3.3: *Altruism is positively related to an individual's intention to report emails that they consider to be phishing.*
H3.4: *Courtesy is positively related to an individual's intention to report emails that they consider to be phishing.*
H3.5: *Emotional Stability is positively related to an individual's intention to report emails that they consider to be phishing.*
H3.6: *Extraversion is positively related to an individual's intention to report emails that they consider to be phishing.*
H3.7: *Self-efficacy is positively related to an individual's intention to report emails that they consider to be phishing.*
H3.8: *Subjective Norms are positively related to an individual's intention to report emails that they consider to be phishing.*

9.4. Methodology

To analyze to which extent human factors affect an individual's intention to report phishing emails and, thus, to test the hypotheses presented in Section 9.3, we conducted an online sur-

vey on Amazon Mechanical Turk (AMT) [25], with a sample size of $n = 284$ participants.[1] The respondents were required to answer a questionnaire to assess their characteristics as well as their cyber security behavior and willingness to report phishing emails. A number of checks were employed to assess the reliability of the data [331]; only responses that passed these checks were considered for hypothesis testing.

9.4.1. Subject Selection

We recruited the participants of our study on AMT. To ensure the quality of the collected data, we required respondents to have a minimum of 1000 previously approved tasks on the platform with an acceptance rate of at least 98%. We also recruited participants only from the US. This choice was made to maintain a high likelihood of having fluent English speakers and to avoid fragmentation in the respondent population. We discuss implications to generalizability of our results in Section 9.6.3.

We follow recent practice underlying the importance of detecting and discarding incorrect responses to maintain a high reliability for studies run on AMT [179, 331]. Following [179], we employ four checks to ensure that unreliable respondents are excluded from data analysis. The four checks are as follows. At the beginning of the survey, participants were required to provide their AMT WorkerID. We removed surveys for which the WorkerID provided by the participant did not match any WorkerID in the list of participants we gathered from the AMT platform for this task, or the same WorkerID occurred multiple times to prevent double entries from a single subject. We also included an attention check question in the survey and only considered the responses of those participants who answered that question correctly. In addition, at the end of the survey, participants had to provide a survey completion code to demonstrate that they have completed the survey. Finally, all participants who completed the survey within 5 minutes[2] were rejected and removed from the experiment. Based on these checks, we discarded 16 participants (5% of respondents); answers from the remaining 284 subjects were included in the analysis.

9.4.2. Survey Design

Our survey aims to assess the human factors influencing an individual's positive cyber security behaviors and intention to report phishing emails. The survey consists of four parts. After a short introduction about the notion of reporting and the purpose of the survey, the participants were asked to provide their demographics, such as their age, education, and current occupation (cf. Table D.5 in Appendix D.3). The second part of the questionnaire comprises questions focusing on the respondents' personality traits and other factors related to their routines at the workplace (i.e., OCB characteristics). The last two sets of questions aim to measure a participant's beliefs, positive cyber security behavior, and intention to report phishing emails, respectively. The survey also includes an open question that allows the

[1] We determine the minimum sample size to obtain a statistical power of 90% by conducting a pilot study involving 100 participants and calculate the final sample size following [224] (full calculations in Appendix D.2). This estimation yielded a required sample size of $n = 267$. Accounting for an estimated 10% of faulty responses, the estimated total sample size for conducting the survey was rounded up to 300 participants. Accordingly, we recruited 200 additional participants for our study.

[2] The expected completion time of the survey is 20 minutes.

participants to share their views regarding why anyone would or not be willing to report suspicious emails. An overview of the survey questions is provided in Appendix D.3 (Table D.6).

To reduce ambiguity and biases in the interpretation of the questions, we took several steps. First, we based the survey items, used to measure the human factors, on existing assessment methods [100, 126, 283, 311] and adapted them to minimize ambiguity and maximize the fit in the context of phishing email reporting. Then, we ran three review rounds. In the first review round, we consulted relevant literature [81, 290] to minimize common pitfalls in questionnaire wording. In the second review round, we administered the questionnaire to six PhD students who provided feedback regarding the clarity of the survey items. Finally, the survey was reviewed by a native English speaker with the specific aim of identifying any remaining ambiguous wording. The gathered feedback was discussed among the authors, and the wording of the questions was finalized accordingly until no further ambiguities or points of improvement emerged.

We used a five-point Likert scale with six items to measure positive cyber security behaviors and four items to measure all the other constructs. Similarly, the attention check question item asks the participants to select a particular choice from the same five-point scale. To minimize ambiguity in the responses, we relied on [290, 344] to choose our wording to define the scale over which users rate their responses. When performing the analysis of the results, the measured factors of each participant are calculated as the average of the answers provided across all items for that specific variable.

9.4.3. Data Analysis

Respondent Demographics

From the answers gathered from the demographic questions in the survey, we first analyzed, by means of their (Pearson) correlation, the relationship of the respondent's demographics with their *positive cyber security behaviors* and *intention to report phishing emails*. This analysis was used to provide descriptive statistics of the survey participants and to provide context for collected respondent data regarding their intention to report suspicious emails.

Reliability and Validity of the Measured Variables

Before assessing the hypotheses, we determined the reliability and validity of the measured items. Following [185] we measure the internal consistency of the measured variables by calculating the Cronbach's α value of the corresponding items, and set the threshold for a satisfactory outcome to 0.7. We also assessed the discriminant validity of the model variables by computing the correlation across the independent and dependent variables to check for multicollinearity problems. Following [139], we assume that there is no multicollinearity if the correlations across all pairs of variables are below the recommended threshold value of 0.8.

Table 9.4: Regression Equations corresponding to the three models

Model	Dependent Variable	Equation
1	Positive Cyber Security Behaviors ($PCSB$)	$PCSB_i = \beta_0 + \beta_1 \cdot Sportsmanship_i + \beta_2 \cdot Conscientiousness_i + \beta_3 \cdot Altruism_i + \beta_4 \cdot Courtesy_i + \beta_5 \cdot EmotionalStability_i + \beta_6 \cdot Extraversion_i$
2	Intention to Report Phishing ($RepInt$)	$RepInt_i = \beta_0 + \beta_1 \cdot PositiveCyberSecurityBehaviors_i$
3	Intention to Report Phishing ($RepInt$)	$RepInt_i = \beta_0 + \beta_1 \cdot Sportsmanship_i + \beta_2 \cdot Conscientiousness_i + \beta_3 \cdot Altruism_i + \beta_4 \cdot Courtesy_i + \beta_5 \cdot EmotionalStability_i + \beta_6 \cdot Extraversion_i + \beta_7 \cdot SubjectiveNorms_i + \beta_8 \cdot SelfEfficacy_i$

Notes: To aid comparison, we report standardized β coefficients. Hence β_0 is centered at 0 for all models. Error terms not reported for brevity.

Hypothesis Evaluation

To evaluate the hypotheses presented in Figure 9.1, we devised three regression models, one for each group of hypotheses. Model 1 encompasses the hypotheses presented in Section 9.3.1 and aims to assess the effect of the OCB characteristics and personality traits on positive cyber security behaviors. Model 2 addresses the hypothesis of Section 9.3.2 and aims to assess the influence of positive cyber security behaviors on an individual's intention to report phishing emails. Finally, Model 3 formalizes the hypotheses of Section 9.3.3 and aims to measure the effect of all identified human factors on the intention to report phishing emails. An overview of the regression equations corresponding to the three models is given in Table 9.4.

We performed a linear regression using the Ordinary Least Squares (OLS) Estimation [156]. When presenting the results of the regression analysis in Section 9.5.3, we report the standardized coefficients of the independent variables to compare the relative magnitude and sign of the effects of these independent variables on the dependent variable.

We perform and report the regression analysis on the three models with and without control variables, where the control variable are derived from the demographic information elicited through the survey (cf. Section 9.4.2). This was to determine the extent to which the control variables modulate the effect of the observed independent variables, which may be an index of additional hidden effects in the models. The results of the regression analysis were used for the evaluation of the hypotheses. We consider factors whose regression coefficients have a p-value lower than 0.05 as statistically *significant*. We note that our hypotheses are directional, hence we only reject the respective null hypotheses when both coefficient sign and statistical significance are aligned with the (statistical validity of the) prediction.

9.4.4. Ethical Aspects

This research was executed under ethical approval from our institution's ethical review board under approval number `ERB2020MCS13`. Participants' WorkerIDs were not transmitted in any form to minimize any risks to our survey's participants. The participants were assured that their answers are used for research purposes only. In the design of the questionnaire, we followed ethical practices for response options [290]. Additionally, in line with the US federal minimum wage of $7.25 per hour [343], each participant that delivered a valid survey response received a compensation of $2.7. With an average completion time of 21 minutes, this equates to an hourly compensation of $7.7.

9.5. Results

After conducting the pilot and the main survey, the total dataset consisted of 284 valid responses. In this section, we first present the descriptive statistics of the respondents and control correlations. Then, we report on the results of the factor reliability and the linear regressions used for testing the hypotheses presented in Section 9.3.

9.5.1. Respondent Demographics

Table 9.5 presents an overview of the demographic information of the 284 participants to our survey. For each control variable, the table reports the frequency and percentage of the participants' answers. We can observe that a similar proportion of male and female participants were part of the survey. Moreover, our sample comprises approximately 60% of adults between 31-50 years of age and most have at least a college education. In line with this profile, 50% of the respondents report being in senior (i.e., not entry-level) job positions. A perhaps surprising statistic emerging from the sample is the seemingly high (45%) fraction of respondents that indicate having fallen for a phishing email. A possible explanation is that the type of task focused on phishing in organizations attracted users with previous experience on the topic. We comment on possible implications for external validity in Section 9.6.3.

Table 9.6 reports an overview of the linear relations between the control variables and an individual's *positive cyber security behaviors* (top rows) and *intention to report phishing emails* (bottom rows). Three of the eight selected controls, namely *Education*, *Current employment duration*, and *Reporting frequency*, showed a significant relationship with *positive cyber security behaviors*. These results indicate that higher educated individuals, being part of the organization for a longer period of time, and who consistently report suspicious emails, are also more inclined to perform actions that benefit the cyber security of the organization. Additionally, *Current employment duration*, *Phishing Victim*, and *Reporting frequency* show a significant positive relationship with an individual's *intention to report phishing emails*. This suggests that individuals that are part of the organization for a longer period of time, who fell for phishing emails in the past, and who consistently report suspicious emails, are also more inclined to report phishing emails.

9.5.2. Questionnaire Reliability and Validity Checks

Reliability

Table 9.7 shows the internal consistency among the variables measured in our questionnaire. We can observe that Sportsmanship, Altruism, Emotional Stability, Self-efficacy, and Positive Cyber Security Behaviors exceeded the recommended threshold value of 0.7, indicating that the constraint on the internal consistency of these measurements is satisfied. On the other hand, for Extraversion and Subjective Norms we dropped survey items E2, and SN3 respectively (cf. Table D.6 in Appendix D.3), to increase the Cronbach's α value to a value close to the recommended threshold. For the remaining variables, namely Conscientiousness, Courtesy and intention to report phishing emails, dropping one or more survey items did not have effect on the increase of the internal consistency of the model variables. However, the results show that the Cronbach's α values are always very close to the threshold, leading to an overall satisfactory internal consistency of our measures.

Validity

Table 9.8 reports the correlations between variables. We can observe that the value of the correlation across all pairs of variables is generally low and below the recommended value of 0.8, indicating no problematic multicollinearity between the considered variables.

Table 9.5: Profile of survey participants

Control	Answer	Freq.	Perc.
Gender (C1)	Male	144	50.7
	Female	140	49.3
	Prefer not to say	0	0.0
	Other	0	0.0
Age (C2)	18–30	68	23.9
	31–50	180	63.4
	> 50	35	12.3
	Prefer not to say	1	0.4
Education (C3)	Primary School	2	0.7
	Secondary/High School	52	18.3
	College/University	230	81.0
Current occupation (C4)	Student	1	0.4
	Employed/Self-employed	271	95.4
	Not employed	9	3.2
	Retired	3	1.1
	Other	0	0.0
Current employee position (C5)	Intern	1	0.4
	Entry-level/Associate	112	39.4
	Manager/Senior manager	147	51.8
	C-level exec./Director/Owner	16	5.6
	Other	8	2.8
Current employee duration (C6)	< half a year	10	3.5
	Between 1/2 year & 2 years	107	37.7
	> than 2 years	167	58.8
Phishing victim (C7)	Yes	129	45.4
	No	155	54.6
Reporting frequency (C8)	Never	45	15.8
	Rarely	49	17.3
	Occasionally	92	32.4
	Frequently	67	23.6
	Always	31	10.9

9.5.3. Hypothesis Evaluation

We tested the hypotheses presented in Fig. 9.1 using the IBM SPSS Statistics software [157]. Fig. 9.2 presents the results of the regression analysis; coefficients are reported in Table D.7 in Appendix D.4. The figure reports the three models vertically, with plots in the top row containing the standardized coefficients of the human factors without introducing the controls (in blue) and with controls (in red). Plots in the bottom row of Fig. 9.2 illustrate the standardized coefficients of the controls given the three models. We observe a very small difference between the value of the coefficients of the independent variables when the set of controls is considered compared to when the controls are omitted. As a consequence,

Table 9.6: Relation of controls with Positive Cyber Security Behaviors and Intention to Report

	Control	Pearson correlation	p-value
Positive Cyber Security Behaviors	C1 (Gender)	0.022	0.707
	C2 (Age)	0.046	0.442
	C3 (Education)	0.124*	0.037
	C4 (Current occupation)	-0.113	0.057
	C5 (Current employment position)	0.055	0.360
	C6 (Current employment duration)	0.192**	< 0.001
	C7 (Phishing victim)	0.035	0.552
	C8 (Reporting frequency)	0.527**	< 0.001
Intention to Report	C1 (Gender)	-0.074	0.213
	C2 (Age)	0.079	0.182
	C3 (Education)	0.046	0.440
	C4 (Current occupation)	-0.014	0.820
	C5 (Current employment position)	-0.084	0.157
	C6 (Current employment duration)	0.216**	< 0.001
	C7 (Phishing victim)	-0.137*	0.021
	C8 (Reporting frequency)	0.357**	< 0.001

Correlation is significant at: the 0.01 level (2-tailed), **; the 0.05 level (2-tailed), *

Table 9.7: Factor Reliability

Factor	Cronbach's α	Observations
Sportsmanship	**0.894**	
Conscientiousness	0.666	
Altruism	**0.777**	
Courtesy	0.658	
Emotional Stability	**0.802**	
Extraversion	0.510	Dropped E2
Self-efficacy	**0.769**	
Subjective Norms	0.673	Dropped SN3
Positive Cyber Sec. Beh.	**0.791**	
Intention to Report	0.693	

the introduced controls do not significantly affect the results of the assessed human factors. Looking at the Adjusted R^2 coefficients reported for the three models in Table D.7, we observe a noticeable effect of controls (chiefly, 'C8 - reporting frequency') only on Model 1, for which their addition explains an additional 14% of the variance in the data (from 38% to 52%). This is unsurprising as higher reporting frequencies can be expected to reflect in overall positive cyber security behaviors. By contrast, the addition of controls in Model 1 and Model 2 only contribute to explaining, approximately, an additional one and three percent of variance, respectively (M1: from 42% to 45%; M2: from 54% to 55%). This suggests that the main effects in the model are appropriate to explain reporting intentions, and that

Table 9.8: Variable Correlations

Variable	1	2	3	4	5	6	7	8	9	10
1. Sportsmanship	1									
2. Conscientiousness	0.305**	1								
3. Altruism	0.147*	0.554**	1							
4. Courtesy	0.448**	0.632**	0.624**	1						
5. Emotional Stability	0.645**	0.276**	0.164**	0.266**	1					
6. Extraversion	0.177**	0.130*	0.396**	0.161**	0.498**	1				
7. Self-efficacy	0.364**	0.570**	0.552**	0.616**	0.225**	0.193**	1			
8. Subjective Norms	0.147*	0.543**	0.513**	0.556**	0.145*	0.235**	0.520**	1		
9. Positive Cyber Security Behaviors	0.114	0.546**	0.513**	0.463**	0.143*	0.272**	0.541**	0.566**	1	
10. Intention to Report	0.178**	0.544**	0.577**	0 .560**	0.190**	0.206**	0.638**	0.591**	0.651**	1

Correlation is significant at the 0.01 level (2-tailed).**
Correlation is significant at the 0.05 level (2-tailed).*

no large hidden effects are likely to be found within the controls.

Figure 9.2: Coefficients of the observed variables and controls

Table 9.9 reports the hypothesis assessment based on the results of the regression analysis when controls are considered.[3] The standardized β value of each hypothesis represents the value of the variable coefficient assessed in the corresponding model. In terms of human factors, we observe that *Conscientiousness* and *Extraversion* have a positive influence on positive cyber security behaviors, while *Altruism*, *Self-efficacy*, and *Subjective Norms* have a positive effect on the intention to report phishing emails. Additionally, *positive cyber security behaviors* is highly influential on an individual's intention to report phishing emails. Next, we discuss each model individually.

Model 1 (OCB → PCSB)

Model 1 aims to test whether the selected human factors, namely Sportsmanship, Conscientiousness, Altruism, Courtesy, Emotional Stability and Extraversion, positively affect an individual's positive cyber security behaviors, as captured by hypotheses H1.1 to H1.6. These predictions are partially supported, where Conscientiousness is the most powerful human factor ($\beta = 0.359; p < 0.001$), indicating that conscientious individuals tend to engage in behaviors that are beneficial for the security of the organization. Extraversion is another human factor positively influencing an individual' positive cyber security behaviors ($\beta = 0.122; p = 0.032$), showing that extraverted individuals are inclined to perform helpful behaviors to protect the organization and other employees.

Model 2 (PCSB → RepInt)

Model 2 aims to test whether positive cyber security behaviors positively affect an individual's intention to report phishing emails (H2.1). This prediction is supported, where the intention to report phishing emails is strongly determined by positive cyber security behaviors ($\beta = 0.630; p < 0.001$). This indicates that reporting potentially dangerous emails falls within the behaviors that an individual may perform outside of his organizational tasks and duties.

Model 3 (OCB + Beliefs → RepInt)

This model hypothesizes that the selected human factors positively affect an individual's intention to report phishing emails (hypotheses H3.1 to H3.8). These hypotheses are partially supported. The results show that beliefs, namely self-efficacy ($\beta = 0.325; p < 0.001$) and subjective norms ($\beta = 0.207; p < 0.001$), are the factors that mostly influence an individual's intention to report phishing emails. Altruism has also a significant impact on an individual's intention to report phishing emails ($\beta = 0.174; p = 0.006$), showing that altruistic individuals tend to act for the benefit of the organization and their colleagues. Surprisingly, our analysis shows that Sportsmanship has a statistically significant *negative* relationship with the intention to report phishing emails ($\beta = -0.169; p = 0.013$). Individuals with

[3]To further verify the robustness of our findings, we re-run all our models adopting a robust OLS regression (which is robust against violations on OLS assumptions) and compared regressed coefficients with those in output of a standard OLS. We find virtually no difference, neither in magnitude nor direction, between the two sets of estimated coefficients for all models.

Table 9.9: Hypothesis testing (estimations from models including control variables)

Hypothesis	Human Factor	Standardized β	p-value	Assessment
H1.1	Sportsmanship	0.021	0.763	Not supported
H1.2	Conscientiousness	0.359	< 0.001	Supported
H1.3	Altruism	0.096	0.132	Not supported
H1.4	Courtesy	0.088	0.192	Not supported
H1.5	Emotional Stability	-0.075	0.259	Not supported
H1.6	Extraversion	0.122	0.032	Supported
H2.1	Safe Cyber Security Behaviors	0.630	< 0.001	Supported
H3.1	Sportsmanship	-0.169	0.013	Not supported*
H3.2	Conscientiousness	0.089	0.138	Not supported
H3.3	Altruism	0.174	0.006	Supported
H3.4	Courtesy	0.071	0.298	Not supported
H3.5	Emotional Stability	0.125	0.053	Not supported
H3.6	Extraversion	-0.054	0.327	Not supported
H3.7	Self-efficacy	0.325	< 0.001	Supported
H3.8	Subjective Norms	0.207	< 0.001	Supported

**Note:* The study results demonstrate a statistically significant negative relationship with the intention to report.

this trait tend to have a high tolerance for less-than-ideal situations such as receiving suspicious emails. In such situations, they might not take further actions to mitigate potential risks and ignore the email, rather than reporting it.

9.6. Discussion and Implications

This section presents the theoretical implications of our study and research directions for practice with respect to phishing reporting intentions and positive cyber security behaviors, followed by a discussion on the threats to validity of this study.

9.6.1. Implications for Theory

This study provides several theoretical implications that can be used in future research, observing the role of cognitive theory in interpreting human behaviors with respect to positive cyber security behaviors within organizations.

The need for a unified model for phishing reporting behavior

Cyber security behaviors are influenced by both human factors at the individual-level (i.e., OCBI characteristics, personality traits and beliefs) and at organization-level (OCBO characteristics). However, previous work that addressed the two levels have done so only within the OCB characteristics, without considering other influential factors, such as personality traits or beliefs (cf. Table D.1).

Comparing results from our findings with prior research underlines the need to develop a cohesive, complete model of reporting behavior to obtain consistent results and derive

effective practices. For instance, on the individual level, prior work showed the positive relationship between emotional stability and security assurance/compliance behaviors [100, 266, 271]; by contrast, our study shows that when considered together with other factors, this relationship may not be significant for positive cyber security behaviors. On the other hand, previous work on extraversion reports mixed results across various security behaviors [100, 271] whereas we find a positive relationship with positive cyber security behaviors. These discrepancies can be ascribed to the fact that emotional stability and extraversion have been often studied in isolation and that they might be less relevant when a broader portfolio of human factors (including organization-related) is considered.

At the organizational level (OCBO), previous work has generally positively related cyber security behaviors with the OCB construct [100]; by contrast, our results suggest only conscientiousness (OCBO) is consistent with previous results (if considering also OCBI the divergence is more prominent). Moreover, our finding of a negative relation between sportsmanship and intention to report is unexpected: individuals with high sportsmanship, by definition, can be reasonably expected to be more inclined to 'take one for the team' (referred to, e.g., the nuance of reporting phishing emails) [203]. Unexpectedly, we find the opposite might be true. An interpretation is that high sportsmanship individuals might not want to create additional burden to other 'members of the team' (in this case, those responsible to handle the reports) because of the, in their view, relatively minor inconvenience of receiving a phishing email. The not significant relationship of conscientiousness with intention to report is surprising as well, because previous literature overall reports positive relationships of conscientiousness (from the Big-Five traits) with specific security behaviors like detecting phishing emails [142]. One possible explanation is that individuals with high conscientiousness (as per OCB) may not consider reporting to be within their 'duties' or that reporting is still a concept misunderstood by many [179].

Future research can extend the scope of our study by evaluating the (combined) effects of the other human factors discussed in Section 9.2.2 as well as of other external factors (e.g., windows of opportunity [123], culture [319]), or factors that can negatively influence reporting. On this line, the bystander effect, i.e., the expectation that others will do the reporting (see Chapter 8), and the ill-perceived liability of reporting, i.e., the assumption that 'it is the duty of the IT department' to deal with phishing (see Chapter 7), can be valuable avenues for research to extend our model and to build a more comprehensive understanding of reporting behaviors in general.

Generic vs. specific cyber security behaviors

Our evaluation shows a strong positive association between individuals' positive cyber security behaviors and their intention to report phishing emails. This relation indicates that employees who report phishing emails (the *specific* behavior), typically act in accordance with the organization's ISP and exhibit security assurance behaviors (the *generic* behavior). However, our study shows that the underlying human factors driving these behaviors could be different. For instance, our results reveal that sportsmanship and altruism have no strong relationships with the generic positive cyber security behaviors whereas these factors do influence the specific behavior of (intention of) reporting, thus casting doubts on their relation with the generic constructs of SAB and SCB. A possible explanation can be that the latter

encompasses behaviors such as 'using password managers' or 'complying with ISP', which poorly align with sportsmanship and altruism. On the other hand, conscientiousness and extraversion appear to have a strong relationship with positive cyber security behaviors, but this does not translate to the specific behavior of reporting. This suggests that relationships between human traits and *specific* cyber security behaviors do not necessarily translate to *generic* behaviors as one might expect. Therefore, researchers and practitioners should be cautious in applying or generalizing their findings to other types of positive cyber security behaviors.

Future research may investigate the impact that negative security behaviors (i.e., SRB and SDB) [100] may have on the individual's reporting actions. The contrast between positive and negative cyber security behaviors may shift an individual's intention to report, and more specifically, it may alter their perspective on what defines *normal* and *abnormal* behaviors [135]. These security behaviors may reduce the intention to report, while counteracting the effect of the positive cyber security behaviors.

Design of innovative training and awareness programs

Our findings can be used to support the design of innovative training and awareness programs. For example, gamification systems employed in phishing reporting can increase the confidence and motivate individuals to perform beneficial cyber security behaviors [171]. When creating such systems to encourage phishing reporting, human factors shaping cyber security behaviors may serve as instruments for fine-tuning the users' interactions with such a system (e.g., the number of false positives may increase when employees are prompted to reporting hits). Our findings suggest that building the employees' confidence in their capability to report potential attacks and the perceived validation from authoritative sources may also increase the individual's motivation to engage in beneficial cyber security behaviors and comply with the organization's policy.

9.6.2. Research Directions for Practice

Our findings also point at interesting research directions to investigate novel approaches for training and awareness programs that organizations often provide, as well as aiming at improving the organizational culture and the overall security posture of an organization (cf. Section 9.2).

Training

This study provides insights relevant to the design of training practices aimed at improving reporting behavior. For example, our study suggests that high sportsmanship is linked to low intention to report a suspicious email. As high-sportsmanship individuals may tend to avoid creating additional work to others because of their own negative experiences, a training program may want to 'fight back' this effect by explicitly gearing the training towards minimizing the negative effects of reporting (i.e., the filed 'complaint') creates on the organization, and maximizing the relevance of the positive outcome in terms of increased overall security. For example, regular training programs aimed at training phishing detection could be extended to cover the process by which reported emails are handled by dedicated staff and

to provide hard-data on the outcomes of the reporting process. Similarly, feedback mechanisms informing the reporter of the effects of their report may help in curtailing the negative effect measured for high sportsmanship individuals. On a similar line, we find self-efficacy and subjective norms to be also factors positively related to the intention to report phishing emails. These findings indicate, for instance, that training programs should focus on the reporting mechanisms as well, rather than (primarily) on the detection of phishing attacks, and employ special interventions aimed at enhancing employees' self-efficacy in this direction. Additional research in this direction is needed to evaluate the effects that these human factors have on training effectiveness.

Awareness

This study's outcomes can also point to future directions to improve the efficacy of cyber security awareness programs. For example, awareness programs can explicitly acknowledge the contribution of conscientious behaviors to support the organization's security posture and reward diligent individuals to motivate them and inspire others to maintain it. Therefore, incentive programs or approaches can be introduced to encourage employees to exceed the formal expectations of the organization, while increasing their awareness of data security. Similarly, employees' altruistic tendencies can be accounted for in awareness programs to encourage the reporting of phishing attacks; for example, awareness programs could further clarify why such behaviors benefit the organization as a whole, and how they can contribute to protecting colleagues that might not be as skilled in recognizing phishing attacks. These insights could be integrated into awareness programs by different means, for example by targeting 'tailored' programs to specific groups or by explicitly acknowledging the role of the single employee in protecting their peers.

Culture

The organization's collective assumptions, values, and perceptions can be a valuable tool to mold positive cyber security behaviors [38]. With respect to phishing reporting behavior, our findings suggest that fostering an organization's culture to encourage individual initiative (self-efficacy) and promote clear expectations within the work environment (subjective norms) may have beneficial effects on reporting and, more in general, on positive cyber security behaviors [225]. The recognition as a cultural value of a conscientious commitment to cyber security and, thus, to the overall well-being of the organization's collective can boost the motivation of individuals to 'keep up the good work'. Moreover, cherishing individual openness and activeness, often observed in extravert individuals, can be an untapped resource for attack mitigation to improve the security posture of the organization. Similarly, altruistic behaviors may be emphasized when defining the organizational culture, where such behaviors are accepted as the norm, and peer collaboration is promoted. Encouraging these behaviors would also lead to a positive outcome for taking protective actions, such as reporting potential phishing attacks. Employees can act as a collective phishing detection mechanism, even in large organizations, enabling fast detection and thwarting of new phishing campaigns with acceptable operational load [196]. Such a mitigation strategy can be 'embedded' in the organization's security stance by developing a sustainable security culture, for example, as part of the organizational culture itself.

9.6.3. Threats to Validity

Construct Validity. The study evaluates an individual's intention to report and not the actual reporting behavior. As a consequence, the construct addressed in this study may not be sufficient to assess reporting behaviors. However, several theories such as PMT [298] and TPB [8] show there is a close relationship between intention and actual behavior and that measures of intention are widely accepted as indicators for actual behaviors.

Internal Validity. To evaluate the internal validity of the study, we assess the measures used for the theoretical model and the conducted survey. Firstly, the reliability of the model is generally supported by satisfactory Cronbach's α values (cf. Table 9.7). On the other hand, some constructs (chiefly, extraversion, and to a lesser extent conscientiousness, courtesy, subjective norms, and intention to report) do show lower internal consistency levels; however, our attempts to increase consistency by removing items to the questionnaires did not help for those constructs, despite being directly adapted from survey questions adopted in the literature [100, 126, 283, 311]. Future work could address how to design more robust measurements for those constructs. Secondly, the sample size for the survey participants is adequate, as the minimum sample size required to assess the outcome is achieved. Finally, we employed several checks to ensure the validity of the survey responses used in the analysis (cf. Section 9.4.1). The collected data, however, might be affected by other bias. For instance, we based the survey items used to measure human factors on existing assessment methods. These survey items are in the form of agree–disagree questions, which can lead to acquiescence response bias [202]. This bias can influence the survey data, where respondents may have the tendency to agree with the questions presented. Future research might design construct-specific questions to mitigate this type of bias [191].

External Validity. To determine the minimum sample size to achieve an acceptable precision in the analysis, we employed a sample size estimation calculation, as described in Appendix D.2. For our study, we recruited participants located in the US through AMT. While it has been shown that AMT workers in the US are representative of the US population when performing security- and privacy-related tasks [292], our results may not generalize to other populations. On the other hand, the respondent demographics reported in Table 9.5 seem to show relatively high rates of phishing victimization rate and high professional seniority. A possible explanation is that the type of task focused on phishing in organizations attracted users with previous experience on the topic. Given that the reference population for our study consists of employed professionals exposed to phishing emails, we consider our findings applicable to that population. The experiment can be reproduced, with respondents from various countries and different levels of email use experience, to minimize the effect of the selection bias and estimate whether different interpretations of the intention to report phishing emails exist.

9.7. Conclusion

In this study, we investigated the influence of human factors on an individual's intention to report phishing emails. To this end, we developed a theoretical model of human factors and their relations with an individual's positive cyber security behaviors and intention to report

phishing emails. We evaluated the model through an experiment within Amazon Mechanical Turk, where 284 participants answered an online survey. The results show that there exists a strong relationship between an individual's positive cyber security behaviors and their intention to report phishing emails. Moreover, among the studied human factors, we observed that self-efficacy, subjective norms, and altruism positively impact reporting intention. However, not all factors that influence intention to report, such as altruism, are positively related to cyber security behaviors. Furthermore, the results reveal that sportsmanship (characterizing subjects who tend to tolerate less-than-ideal situations, such as the nuance of reporting a phishing email) hinders an individual's intention to report phishing emails.

Our findings provide an answer to the second part of RQIV and shed light on theoretical a practical implications of human factors in relation to phishing reporting. The recommendations provided in this chapter suggest that more research and experiments are required to evaluate how human factors can be leveraged to improve an organization's security posture, for example, in terms of training and awareness programs as well as to foster an organization's security culture.

10

Conclusions

10.1. Summary of Contributions

This thesis aims at filling the gap in understanding social engineering attacks, specifically in terms of tailored phishing and countermeasures against it. We therefore structured this work around the following main research question:

Main RQ: *What are the current gaps in our understanding of tailored phishing attacks from the target, attacker, and defender perspectives, and which technological and organizational methods can be employed to address these gaps?*

To investigate this question, in Part I, we explored theories and models of human cognition involved in SE attacks and developed a framework to evaluate and contextualize research results in SE (Chapter 3). The framework was instrumented to carry out a systematic literature review of empirical SE research that focuses on experimental characteristics and core cognitive features from both attacker and target perspectives (Chapter 4). The literature review, therefore, helped to identify gaps and open research questions in SE research, including gaps in our understanding of SE attacks, such as tailored phishing.

In Part II, we carried out a field experiment simulating a tailored phishing campaign which showed the effects of using target-related information in the phishing emails across organizations and employee categories (Chapter 5). Our study on technological mitigation strategies for phishing attacks yielded a proof-of-concept detection and decision support system to defend against attacks aiming at stealing web credentials (Chapter 6). Further analysis of our simulated phishing campaign led us to explore and propose an organizational mitigation strategy against sophisticated attacks, such as spear and tailored phishing, in the form of an improved phishing reporting process (Chapter 7).

We further examined organizational mitigation strategies in Part III by carrying out two studies that investigate phishing reporting behavior. We interviewed employees of a small IT company to understand their reaction to a simulated phishing campaign (Chapter 8). To understand what factors influence the intention to report phishing at organizations, we ran an online questionnaire to evaluate the relationships of different human factors and intention to report phishing (Chapter 9). The results of these studies revealed a series of impli-

cations for research and practice that extend our comprehension of possible organizational mitigation strategies against advanced phishing attacks.

Following, we revisit the main results and findings contributing to answer each of the research questions derived from the Main RQ.

10.2. Characterizing the human attack surface

The SE domain is related to a variety of disciplines, such as human computer interaction, sociology and psychology, among others. This multidisciplinarity makes it difficult to identify gaps and open research questions and to interpret experimental result. Therefore, we asked:

RQI: *How can we characterize the SE attack surface to evaluate and contextualize research results and identify gaps in empirical SE research?*

This question was addressed in Chapter 3 where we investigated well-established theories of human cognition, and examined the cognitive processes that can be affected by an attacker during an SE attack:

> **The SE attack surface can be characterized by relating attack effects and techniques to specific cognitive features and processes of the targets.** The processes of human perception, attention and elaboration can be conditioned by incoming stimuli and contextual variables to produce a behavior. The mapping of these processes and features to the SE domain, such as the attack medium, attacker assumptions and persuasion techniques, provides a common structure for comparisons across different SE attacks. The resulting framework can be used to analyze real and simulated SE attacks, and serves as an instrument to identify gaps in empirical SE research.

To illustrate how the framework can be used to evaluate and contextualise research results, we applied it to two simulated SE attacks and two real attack cases. The analysis reveals, for example, that the framework forces the identification of relevant attacker and target parameters (i.e., properties characterizing the context in which the attack occurs) that might not be explicitly included in the experiment design of a simulated SE attack. This can be a valuable insight as the (mis)matches between attacker assumptions and target characteristics are a significant explanatory factor in SE susceptibility [132]. Similarly, the framework can aid isolating factors that are difficult to recognize without a reference to the features of human cognition, such as, effects on perception of previous stimuli (e.g., priming [266]) or the attention type expected in the participant (e.g., delivering attack artefacts during high attentional load [242]). From a practical point of view, the forced identification of properties characterizing the context in which the attack occurs, and the match thereof, has the potential to improve risk metrics for different typologies of attacks, for example, based on the level of attack adaptation to the targets [320] or the presence of multiple target-attacker interactions [149]. These observations highlight the potential of the framework to enable a systematic comparison of different SE attacks based on their cognitive features, and to identify gaps in experiments simulating SE attacks.

With the framework of Chapter 3 at hand, we were able to address RQII:

RQII: *What are the open gaps between the features of human cognitive processes and empirical research in SE, including future research directions?*

By reviewing 169 papers in empirical SE research, we identified and characterized the open gaps between the features of human cognitive processes and empirical research. We identified the following gaps in Chapter 4:

Gap between real attacks and attacks simulated in the studies: most experiments only partially reflect the complexity of real SE attacks and investigate only a small portion of the overall attack space. For instance, the majority of studies focused on one-step one-stimulus attack scenarios (e.g., email and click on link), as opposed to more sophisticated and increasingly more relevant – multi-step multi-modal attacks where attack interactions may cross multiple media, applications, and devices (e.g., social network to phishing website to download).

The SE attack surface is vast: the exploitable SE attack surface appears much larger than the coverage provided by the current body of research where, for example, despite their high relevance for both attack design and defense, factors such as targets' context (e.g., device type, task and social context) and cognitive processes (e.g., attention type, triggers of anomalies) are often ignored or not explicitly considered in experimental designs.

Studies are focused on a few experimental setups only: the literature tends to employ certain experimental methods with specific populations whereby the most commonly investigated scenarios consist of lab experiments with a generic population (e.g., crowd-sourced online questionnaires) and field experiments at organizations (e.g., embedded phishing exercises). This can make the obtained results of limited explanatory power.

Inconsistent constructs of experimental outcomes: experimental constructs devised to measure attack success rate vary, from clicking a link, opening an attachment to visiting a website. Each of these constructs arguably measures different degrees of attack success (which, depending on the threat model, do not necessarily lead to a security impact) and, conversely, may lead to conflicting findings.

Lack of common reference for targetization: the effects of different pretexts and varied targetization levels (i.e., to what extent an attack was adapted to the recipient) are overall marginally considered. Attack targetization is responsible for large changes in expected success rates, however we do not know to what extent tailored phishing techniques improve the attack success rate or when different pretexts work best.

Future experiments in SE can address richer, more complex scenarios across different domains, for example, with multi-step multi-modal simulations that go beyond email-to-click, such as credential submissions with multi-factor authentication [30] or various QR code delivery methods [354]. Contextual variables and effects on cognitive features are especially difficult to control or measure. Supplementary techniques from the fields of cognitive science and (social) psychology can provide methodological insights on how to investigate factors such as, the role of perception [90], attention [207] or elaboration [44, 153]. Further

limitations in empirical SE research may be mitigated with more realistic laboratory experiments (e.g., realistic user interfaces and complete attack process [273, 322]) and in vivo observational studies (e.g., in collaboration with service providers [391]). Qualitative insights of confounding factors (otherwise difficult to measure quantitatively) can shed light on user context (e.g., briefly after the embedded training [93]) and on targetization effects (e.g., characterizing the premise alignment [323]).

10.3. Tailored phishing and potential counter-strategies

Among the gaps identified in Chapter 4, generic, un-targeted phishing constitutes the overwhelming majority of experiments; this leads to a limited comprehension of the effectiveness of phishing attacks tailored to their targets. As tailored phishing is becoming increasingly relevant to the overall threat landscape, with potentially high impact and relatively low effort from the attacker, we asked the question:

RQIII: *How effective are tailored phishing campaigns in deceiving targets to perform an action, and what strategies can be employed to mitigate these attacks?*

By performing two simulated tailored phishing campaigns against a Dutch university and a consultancy company, in Chapter 5 we learned that:

> **Tailored phishing attacks can achieve a high success rate** (in terms of credential submissions) if compared to the yield of an average phishing campaign. Overall, employees are highly susceptible, with attack success rates between 10% and 30% across user roles and organizations.
>
> **The effectiveness of tailored phishing depends on the target environment and user role.** In our experiment, company employees, as opposed to university employees, are significantly more susceptible to our attack, with junior employees the most vulnerable category in both environments.
>
> **Most users that will fall for the attack are likely do so in the very few hours after attack delivery:** 50% of submissions occurred within the first 2 hours from attack delivery and more than 75% of submissions occurred within 4 hours.

By investigating the effects of persuasion techniques and notification methods on the attack success rate, we found that:

> **There is no significant effect for the introduction of persuasion techniques in our tailored phishing campaigns.** The means by which persuasion techniques are implemented has a sizeable effect, although it is not stable across experiment conditions (i.e., different notification methods and target characteristics lead to different outcomes).

This suggests that 'baseline' persuasion techniques (commonly employed in generic campaigns) can be superseded by the overall 'persuasive' effect of a well tailored phishing email. The adoption of notification methods (the means by which persuasion techniques are implemented) appears, instead, more important than the mere presence of a persuasion technique in a tailored attack. For example, the effect of specific persuasion techniques (e.g., Scarcity and Authority) may be enhanced by notification methods which 'move' the cognitive attack

from the body text of the phishing email to a more prominent position, such as the subject or signature. However, the effectiveness of these attack techniques (combination of persuasion techniques and notification methods) can vary significantly across organization types (i.e., industry vs. academia) and professional roles (i.e., junior, support and senior roles).

Our findings suggest that current user training and awareness programs aimed at immunizing individuals against widely used persuasion techniques (Scarcity, Authority, etc.) may be off-target in a highly-tailored phishing scenario. In this context, users' lack of knowledge on internal organization processes (i.e., junior staff or newly hired personnel) is being exploited by an attacker to build credible pretexts. Therefore, specific training targeted towards this baseline of vulnerable users may help to reduce the attack surface [363], while an efficient mechanism for user reporting (especially with experienced users) may mitigate the impact of the remaining fraction of users that fell for the attack early on [56]. Overall, more replication studies evaluating the effects of target-related information in phishing, against different populations and organizational settings, are necessary to improve our understanding of tailored phishing and help devise more effective mitigation strategies.

Chapter 6 addresses the second half of RQIII by exploring a technological mitigation strategy against phishing websites that are often the payload of phishing campaigns, such as the campaign in Chapter 5. Our proposition to mitigate web-based phishing attacks consist of an experimental tool with the following characteristics:

> **Integrated phishing detection techniques** (i.e., back-end detection logic) **and HCI ingredients** (i.e., front-end user notification methods) to evaluate, characterize, and refine the interaction between phishing decision support, and the final user.
>
> **'zero-hour' phishing detection capability** by relying on search engines to identify which website a phishing page is replicating by means of textual and visual features extracted from an unknown page. This removes the reliance on predefined corpora of brand representations (e.g., URLs, screenshots).
>
> **Various warning methods for user notification**, such as blocking warnings on successful detection and 'retrospective' notifications of past phishing encounters. The front-end is packaged into a browser extension to facilitate the deployment of new experiments with varying numbers of participants.

The use of visual features, in addition to textual features, allows achieving a phishing detection accuracy of 99.66% (vs. 98.98% previous work) on phishing data from 2019. Whereas the improved detection performance compared to previous similar work is small in absolute terms, it represents a ≈ 70% reduction in error rate. This is a significant advance in web page analysis-based techniques, as with high numbers of websites to check even small error rates can get in the way of user reliance on decision-support tools. A key feature of the proposed approach is its 'zero-hour' detection capability: using search engines to identify legitimate websites removes the need to rely on corpora of predetermined URLs, screenshots or training sets, which is a major shortcoming of current visual similarity-based detection methods. However, this comes at the price of longer runtimes that may severely affect the tool's usability. This limitation is mitigated by leveraging various risk communication methods that do not disrupt the user experience, but allow the user to remediate previous 'bad' decisions,

analogous to password breach alerts in modern browsers [334]. Future work can assess the efficacy of tool's risk advice, as well as improve runtime performances, for example, by evaluating tools usability in real-world conditions or by optimizing the caching system to reduce the number of requests.

Whereas technological solutions can help with certain attack classes, such as phishing websites of Chapter 6, it largely remains difficult to defend against sophisticated phishing. In Chapter 7, we follow up on the mitigation strategies of RQIII by promoting more effective *response* procedures in organizations to counter tailored phishing. Preliminary evidence suggests that some users are 'phishing champions', i.e. naturally predisposed to identify anomalies between the communication processes employed by tailored phishing attacks and the 'normal' ones employed by an organization. Therefore, we propose to:

> **Leverage the natural 'immunity' of (some) employees in an organization**, such as employees with a deep knowledge of the 'normal' processes within the organization and a natural ability to detect 'anomalies' in related communication, to mitigate tailored attacks.

However, only a few users typically report phishing emails, and the rationale and factors behind this are poorly explored in the literature. The preliminary results uncover the inability of university employees (that reported to IT the attack of Chapter 5) to generalize the rationale for notifying a suspicious email, as opposed to their consistent rationale in classifying a phishing email as such. For example, the reporting procedure is ill-perceived in terms of effort and liability (i.e., it is someone else duty to deal with security incidents). On the other hand, a higher sense of responsibility to protect colleagues or their organization motivate employees to act. Similarly, the awareness of perceived sophistication of the attack, and the uncertainty on their decision to act emerge as underlying factors for reporting. Therefore, to enable the proposed mitigation strategy, we need to:

> **Increase reporting incidence** by, for example, gearing awareness campaigns towards reporting phishing and specific training at 'phishing champions'.
>
> **Assess the quality of reports** by, for instance, developing reputation-based methods to assign risk scores to specific reports, such as those from 'phishing champions'.

Having a reliable reporting process and a defined risk metric does not address the attack velocity issue highlighted in Chapter 5. It remains crucial to operationalize the risk metrics to anticipate the containment phase as soon as possible after the first few user notifications. Our findings suggest that reports may indeed arrive 'soon enough' to enable this strategy. To operationalize this idea, researchers can investigate an automatic response procedure that can be initiated when sufficiently many high-risk reports are collected, such as attempted in [196].

10.4. Why do people report phishing

Following on the findings of Chapter 7, we understand that a more efficient phishing reporting process based on employees better predisposed to detect complex attacks and eager to re-

port can aid the resilience of the organization as a whole. However more research into users' reasoning and factors that influence their decisions is needed. Therefore, we asked RQIV:

RQIV: *What rationale do users follow when deciding to report a phishing attack and what influences their decisions?*

In Chapter 8, the analysis of reactions of employees at a small IT company after a tailored phishing attack helps answering RQIV, and reveals that the participants' rationale for reporting our phishing attack includes:

> **The protection of other colleagues** from the consequences of the attack and **the expectation that sufficiently many reports may protect the whole group**.

The reasons to *not* report the attack concern:

> **The lack of knowledge on how to report** a phishing email, the belief that **someone else should do it** (delegation to others) and that **reporting will not have any effect**.

Furthermore, user reporting of our phishing attack was affected by:

> **The mismatch of expectations** (i.e., the detection of an inconsistent pattern) as the primary method for detecting the attack, possibly due to the limited size of the company where 'everyone knows everyone'.
>
> **The tailored nature of the attack** prompting certain employees to investigate further, such as carrying out 'whois' look-ups or submitting fake credentials, and eventually notifying the wider group.

The observed collective defense may be attributed to the network effects inherent in small organizations, which might mitigate the bystander effect that is frequently observed in larger organizational settings. This suggests that it can be challenging for an attacker to carry out tailored phishing campaigns in an environment where members know and help each other, in contrast with 'siloed' structures typical of larger enterprises. Future research may investigate how larger organizations can achieve similar effects, in terms of 'group defense', against sophisticated phishing attacks, for instance, with improved communication between users who report and the department collecting such reports [195]. Similarly to the findings of Chapter 7, the participants' mentions of reasons to not report signal that there may be a need for security awareness programs to emphasize the underlying reasons why reporting is important, for example, by presenting real case studies that employees can relate to [204].

By developing and testing a theoretical model that explains intention to report phishing in organizations, in Chapter 9, we answer the second part of RQIV showing that different personality traits, beliefs, attitudes towards the organization and co-workers affect individuals' intention to report phishing emails and other positive cyber security behaviors, such as using password managers or adhering to the Infromation Securiy Policy (ISP). Specifically:

> **Self-efficacy, subjective norms, and altruism positively impact reporting intention**; while a high **sportsmanship** attitude (the tendency to tolerate less-than-ideal situations) **hinders an individual's intention to report phishing** emails.

> Employees who report phishing emails, typically act in accordance with the organization's ISP and exhibit other positive security behaviors. However, the underlying human factors driving these behaviors could be different: **relationships between human traits and specific cyber security behaviors**, such as reporting phishing, **do not necessarily translate to generic behaviors**, such as complying with the ISP.

Our findings can support researchers and practitioners in the design of better training practices and awareness programs and help organizations to create a security culture. For instance, our results on high-sportsmanship individuals, who usually tend to avoid filing complaints, might be explained by the desire to avoid creating additional burden to other 'members of the team'. Building on the findings of Chapter 7 and 8, a training program may mitigate this effect by stressing the relevance of phishing reporting in terms of increased overall security. Moreover, employees' altruistic tendencies can be accounted for in awareness programs by further clarifying why such behaviors benefit their not-as-skilled colleagues in recognizing phishing attacks.

Comparing our findings with prior research casts doubts on previous research results, potentially due to the inclusion of additional factors, such as organization-related factors. This underlines the need to develop a cohesive, complete model of reporting behavior that includes a broader portfolio of human factors and specific security behaviors. For instance, conscientiousness and extraversion appear to have a strong relationship with generic cyber security behaviors, but this does not translate to the specific behavior of reporting; to the contrary, we observed sportsmanship and altruism to influence reporting. Future research can extend the scope of our study by evaluating the (combined) effects of the other human and external factors.

10.5. Final remarks

As the well-being of our society is strongly intertwined with the functioning of digital communications, this thesis advances SE research by identifying, estimating and mitigating the associated risks. Given that the impact of SE attacks is affected by attack features, target characteristics and the countermeasures in place, we account for all three perspectives in our efforts. We explore the target perspective in SE by reviewing the interplay between cognitive effects and SE attack features. We delve into the attacker perspective by analyzing the effects of tailored phishing attacks and, finally, we study the defender perspective by investigating innovative phishing mitigation strategies.

In all three perspectives, our reasoning revolves around the importance of considering users as the main action point to understand and improve the security of computer systems. We thus hope to show in this thesis that overlooking users or blaming them for security failures – 'humans as the weakest link' – is counterproductive, if not outdated. Instead, we should focus our efforts to develop more resilient systems where the user can be an asset – where the 'weakest link' is allowed fail, but gracefully.

References

[1] **Abbasi, A., Mariam Zahedi, F., and Chen, Y.** Phishing susceptibility: The good, the bad, and the ugly. In *International Conference on Intelligence and Security Informatics* (2016), IEEE, pp. 169–174.

[2] **Abdelhamid, N., Ayesh, A., and Thabtah, F.** Phishing detection based Associative Classification data mining. *Expert Systems with Applications 41*, 13 (Oct. 2014), 5948–5959.

[3] **Abdelnabi, S., Krombholz, K., and Fritz, M.** VisualPhishNet: Zero-Day Phishing Website Detection by Visual Similarity. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security* (2020), CCS '20, ACM, pp. 1681–1698.

[4] **Aburrous, M., Hossain, M., Dahal, K., and Thabtah, F.** Experimental Case Studies for Investigating E-Banking Phishing Techniques and Attack Strategies. *Cognitive Computation 2*, 3 (2010), 242–253.

[5] **Adebowale, M., Lwin, K., Sánchez, E., and Hossain, M.** Intelligent web-phishing detection and protection scheme using integrated features of Images, frames and text. *Expert Systems with Applications 115* (2019), 300–313.

[6] **Afroz, S., and Greenstadt, R.** PhishZoo: Detecting Phishing Websites by Looking at Them. In *Int. Conference on Semantic Computing* (2011), IEEE, pp. 368–375.

[7] **Agrafiotis, I., Nurse, J. R., Buckley, O., Legg, P., Creese, S., and Goldsmith, M.** Identifying attack patterns for insider threat detection. *Computer Fraud & Security 2015*, 7 (July 2015), 9–17.

[8] **Ajzen, I.** The theory of planned behavior. *Organizational Behavior and Human Decision Processes 50*, 2 (1991), 179–211.

[9] **Akbar, N.** Analysing persuasion principles in phishing emails. Master's thesis, University of Twente, 2014.

[10] **Akhawe, D., and Felt, A. P.** Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness. In *USENIX Security Symposium* (2013), USENIX Association, pp. 257–272.

[11] **Aleroud, A., and Zhou, L.** Phishing environments, techniques, and countermeasures: A survey. *Computers & Security 68* (2017), 160 – 196.

[12] **Algarni, A., Xu, Y., and Chan, T.** An empirical study on the susceptibility to social engineering in social networking sites: The case of Facebook. *European Journal of Information Systems 26*, 6 (2017), 661–687.

[13] **Alkhalil, Z., Hewage, C., Nawaf, L., and Khan, I.** Phishing Attacks: A Recent Comprehensive Study and a New Anatomy. *Frontiers in Computer Science 3* (2021), 6.

[14] **Allodi, L.** Economic Factors of Vulnerability Trade and Exploitation. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (2017), CCS '17, ACM, pp. 1483–1499.

[15] **Allodi, L., Chotza, T., Panina, E., and Zannone, N.** The Need for New Antiphishing Measures Against Spear-Phishing Attacks. *IEEE Security & Privacy 18*, 2 (2020), 23–34.

[16] **Allodi, L., and Massacci, F.** Comparing Vulnerability Severity and Exploits Using Case-Control Studies. *ACM Transactions on Information and System Security 17*, 1 (2014), 1:1–1:20.

[17] **Alnajim, A., and Munro, M.** An Anti-Phishing Approach that Uses Training Intervention for Phishing Websites Detection. In *International Conference on Information Technology: New Generations* (2009), IEEE, pp. 405–410.

[18] **Alseadoon, I., Chan, T., Foo, E., and Nieto, J.** Who is more susceptible to phishing emails?: A Saudi Arabian study. In *Australasian Conference on Information Systems* (2012), AISEL.

[19] **Alseadoon, I., Othman, M., Foo, E., and Chan, T.** Typology of phishing email victims based on their behavioural response. In *Americas Conference on Information Systems* (2013), vol. 5, AISEL, pp. 3716–3724.

[20] **Alseadoon, I., Othman, M. F. I., and Chan, T.** What Is the Influence of Users' Characteristics on Their Ability to Detect Phishing Emails? In *Advanced Computer and Communication Engineering Technology* (2015), Lecture Notes in Electrical Engineering, Springer, pp. 949–962.

[21] **Altena, L.** Exploring effective notification mechanisms for infected IoT devices. Master's thesis, TU Delft, 2018.

[22] **Althobaiti, K., Jenkins, A. D. G., and Vaniea, K.** A Case Study of Phishing Incident Response in an Educational Organization. *Proceedings of the ACM on Human-Computer Interaction 5*, CSCW2 (2021), 338:1–338:32.

[23] **Amabile, T., Schatzel, E., Moneta, G., and Kramer, S.** Leader Behaviors and the Work Environment for Creativity: Perceived Leader Support. *The Leadership Quarterly 17* (2004), 5–32.

[24] **Amazon**. Alexa - Top sites. `https://web.archive.org/web/20220101025437/https://www.alexa.com/topsites`, 2021. Accessed: 2022-04-09.

[25] **Amazon**. Amazon Mechanical Turk. `https://www.mturk.com/`, 2022. Accessed: 2023-10-22.

[26] **Anderson, B., Vance, A., Kirwan, C., Eargle, D., and Jenkins, J.** How users perceive and respond to security messages: A NeuroIS research agenda and empirical study. *European Journal of Information Systems 25*, 4 (2016), 364–390.

[27] **Anderson, J. R.** *Cognitive psychology and its implications*. Worth publishers, 2000.

[28] **APWG**. Trends Report Q1 2020. Tech. rep., Anti-Phishing Working Group, 2020.

[29] **Arstecnica**. Lapsus$ and SolarWinds hackers both use the same old trick to bypass MFA. `https://edu.nl/qdwj8`, 2022. Accessed: 2023-10-22.

[30] **Arstecnica**. Still using authenticators for MFA? Software for sale can hack you anyway. `https://edu.nl/y3rjh`, 2023. Accessed: 2023-10-22.

[31] **Arthur Jr, W., Bennett Jr, W., Stanush, P. L., and McNelly, T. L.** Factors That Influence Skill Decay and Retention: A Quantitative Review and Analysis. *Human Performance 11*, 1 (1998), 57–101.

[32] **Baars, B.** The conscious access hypothesis: Origins and recent evidence. *Trends in Cognitive Sciences 6*, 1 (2002), 47–52.

[33] **Baars, B., and Franklin, S.** How conscious experience and working memory interact. *Trends in Cognitive Sciences 7*, 4 (2003), 166–172.

[34] **Baars, B. J.** *In the Theater of Consciousness: The Workspace of the Mind*. Oxford University Press, 1997.

[35] **Bada, M., Sasse, A. M., and Nurse, J. R. C.** Cyber Security Awareness Campaigns: Why do they fail to change behaviour? *CoRR arXiv:1901.02672* (2019).

[36] **Baddeley, A.** The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences 4*, 11 (2000), 417–423.

[37] **Ball, L., Ewan, G., and Coull, N.** Undermining: social engineering using open source intelligence gathering. In *International Conference on Knowledge Discovery and Information Retrieval* (2012), Scitepress Digital Library, pp. 275–280.

[38] **Bansal, G.** Got Phished! Role of Top Management Support in Creating Phishing Safe Organizations. *MWAIS 2018 Proceedings* (2018).

[39] **Barracuda**. Spear-phishing Report 2021. `https://www.barracuda.com/reports/spear-phishing-report-6`, 2021. Accessed: 2021-12-14.

[40] **Barracuda**. Spear-phishing Report 2022. `https://www.barracuda.com/reports/spear-phishing-report-7`, 2022. Accessed: 2023-10-14.

[41] **Barrick, M. R., and Mount, M. K.** The Big Five Personality Dimensions and Job Performance\: A Meta-Analysis. *Personnel Psychology 44*, 1 (1991), 1–26.

[42] **Baryshevtsev, M., and McGlynn, J.** Persuasive Appeals Predict Credibility Judgments of Phishing Messages. *Cyberpsychology, Behavior, and Social Networking 23*, 5 (2020), 297–302.

[43] **Bellingcat**. FSB Team of Chemical Weapon Experts Implicated in Alexey Navalny Novichok Poisoning. `https://edu.nl/vaxy9`, 2020. Accessed: 2023-10-22.

[44] **Bellur, S., and Sundar, S. S.** How Can We Tell When a Heuristic Has Been Used? Design and Analysis Strategies for Capturing the Operation of Heuristics. *Communication Methods and Measures 8*, 2 (Apr. 2014), 116–137. Publisher: Routledge.

[45] **Benenson, Z., Gassmann, F., and Landwirth, R.** Unpacking Spear Phishing Susceptibility. In *Financial Cryptography and Data Security* (Cham, 2017), M. Brenner, K. Rohloff, J. Bonneau, A. Miller, P. Y. Ryan, V. Teague, A. Bracciali, M. Sala, F. Pintore, and M. Jakobsson, Eds., LNCS, Springer, pp. 610–627.

[46] **Bird, C.** Interviews. In *Perspectives on Data Science for Software Engineering*. Morgan Kaufmann, 2016, pp. 125–131.

[47] **Blond, S. L., Uritesc, A., Gilbert, C., Chua, Z. L., Saxena, P., and Kirda, E.** A Look at Targeted Attacks Through the Lense of an {NGO}. In *USENIX Security Symposium* (2014), USENIX Association, pp. 543–558.

[48] **Bowen, B. M., Devarajan, R., and Stolfo, S.** Measuring the human factor of cyber security. In *2011 IEEE International Conference on Technologies for Homeland Security (HST)* (Nov. 2011), IEEE, pp. 230–235.

[49] **Bradski, G., and Kaehler, A.** *Learning OpenCV*. O'Reilly Media, Inc., 2008.

[50] **Braun, V., and Clarke, V.** Using thematic analysis in psychology. *Qualitative Research in Psychology 3*, 2 (2006), 77–101.

[51] **Bullee, J., Montoya, L., Junger, M., and Hartel, P.** Telephone-based social engineering attacks: An experiment testing the success and time decay of an intervention. In *Inaugural Singapore Cyber Security R&D Conference* (2016), IOS Press, pp. 107–114.

[52] **Bullee, J.-W.** *Experimental social engineering: investigation and prevention*. PhD Thesis, Centre for Telematics and Information Technology (CTIT), 2017.

[53] **Bullee, J.-W., and Junger, M.** How effective are social engineering interventions? A meta-analysis. *Information and Computer Security 28*, 5 (2020), 801–830. Publisher: Emerald Publishing Limited.

[54] **Bullee, J.-W., Montoya, L., Junger, M., and Hartel, P.** Spear phishing in organisations explained. *Information & Computer Security 25*, 5 (Jan. 2017), 593–613. Publisher: Emerald Publishing Limited.

[55] **Bullée, J.-W. H., Montoya, L., Pieters, W., Junger, M., and Hartel, P. H.** The persuasion and security awareness experiment: reducing the success of social engineering attacks. *Journal of Experimental Criminology 11*, 1 (Mar. 2015), 97–115.

[56] **Burda, P., Allodi, L., and Zannone, N.** Don't Forget the Human: a Crowdsourced Approach to Automate Response and Containment Against Spear Phishing Attacks. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)* (Sept. 2020), IEEE, pp. 471–476.

[57] **Burda, P., Allodi, L., and Zannone, N.** Dissecting Social Engineering Attacks Through the Lenses of Cognition. In *European Symposium on Security and Privacy Workshops* (Sept. 2021), IEEE, pp. 149–160.

[58] **Burda, P., Allodi, L., and Zannone, N.** A Decision-Support Tool for Experimentation on Zero-Hour Phishing Detection. In *Foundations and Practice of Security* (2022), G.-V. Jourdan, L. Mounier, C. Adams, F. Sèdes, and J. Garcia-Alfaro, Eds., vol. 13877 of *LNCS*, Springer Nature Switzerland, pp. 443–452.

[59] **Burda, P., Allodi, L., and Zannone, N.** Cognition in Social Engineering Empirical Research: a Systematic Literature Review. *ACM Transactions on Computer-Human Interaction* (Nov. 2023). Just Accepted.

[60] **Burda, P., Altawekji, A. M., Allodi, L., and Zannone, N.** The Peculiar Case of Tailored Phishing against SMEs: Detection and Collective Defense Mechanisms at a Small IT Company. In *2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)* (2023), pp. 232–243.

[61] **Burda, P., Chotza, T., Allodi, L., and Zannone, N.** Testing the Effectiveness of Tailored Phishing Techniques in Industry and Academia: A Field Experiment. In *Proceedings of the 15th International Conference on Availability, Reliability and Security* (New York, NY, USA, Sept. 2020), ARES '20, ACM, pp. 1–10.

[62] **Burns, A., Johnson, M. E., and Caputo, D. D.** Spear phishing in a barrel: Insights from a targeted phishing campaign. *Journal of Organizational Computing and Electronic Commerce 29*, 1 (2019), 24–39. Publisher: Taylor & Francis.

[63] **Butavicius, M., Parsons, K., Pattinson, M., and McCormac, A.** Breaching the Human Firewall: Social engineering in Phishing and Spear-Phishing Emails. In *ACIS 2015 Proceedings* (2015), vol. 98, AISEL, p. 11.

[64] **Bórquez, J.** Convert any image to pure CSS. `https://javier.xyz/img2css/`, 2020. Accessed: 2020-10-26.

[65] **Canfield, C., Fischhoff, B., and Davis, A.** Quantifying Phishing Susceptibility for Detection and Behavior Decisions. *Human Factors 58*, 8 (2016), 1158–1172.

[66] **Caputo, D., Pfleeger, S., Freeman, J., and Johnson, M.** Going spear phishing: Exploring embedded training and awareness. *IEEE Security and Privacy 12*, 1 (2014), 28–38.

[67] **Cavacini, A.** What is the best database for computer science journal articles? *Scientometrics 102*, 3 (2015), 2059–2071.

[68] **Cetin, O., Hanif Jhaveri, M., Gañán, C., van Eeten, M., and Moore, T.** Understanding the role of sender reputation in abuse reporting and cleanup. *Journal of Cybersecurity 2*, 1 (2016), 83–98.

[69] **Chen, H., Beaudoin, C., and Hong, T.** Securing online privacy: An empirical test on Internet scam victimization, online privacy concerns, and privacy protection behaviors. *Computers in Human Behavior 70* (2017), 291–302.

[70] **Chen, Y., Zahedi, F. M., Abbasi, A., and Dobolyi, D.** Trust calibration of automated security IT artifacts: A multi-domain study of phishing-website detection tools. *Information & Management 58*, 1 (2021), 103394.

[71] **Chiew, K., Chang, E., Sze, S., and Tiong, W.** Utilisation of website logo for phishing detection. *Computers & Security 54* (2015), 16–26.

[72] **Chiew, K., Choo, J., Sze, S., and Yong, K.** Leverage Website Favicon to Detect Phishing Websites. *Security and Communication Networks* (2018), 1–11.

[73] **Cialdini, R.** *Influence: The Psychology of Persuasion*. Harper Business, 1984.

[74] **Cialdini, R.** *Influence: The Psychology of Persuasion*. HarperCollins, 2009.

[75] **Cialdini, R.** *Pre-suasion: A revolutionary way to influence and persuade*. Simon and Schuster, 2016.

[76] **Cialdini, R. B., and Goldstein, N. J.** The science and practice of persuasion. *Cornell Hotel and Restaurant Administration Quarterly 43*, 2 (2002), 40–50. Publisher: Sage Publications Sage CA: Thousand Oaks, CA.

[77] **Cidon, A., Gavish, L., Bleier, I., Korshun, N., Schweighauser, M., and Tsitkin, A.** High Precision Detection of Business Email Compromise. In *USENIX Security Symposium* (2019), USENIX Association, pp. 1291–1307.

[78] **City of Austin**. Fraudulent QR codes found on Austin parking pay stations - AustinTexas.gov. `https://edu.nl/xpev8`, 2022. Accessed: 2023-10-22.

[79] **CleverTap**. Data-Backed Secrets to Successful Push Notifications 2018 Report. `https://clevertap.com/blog/2018-push-notification-report/`, 2018. Accessed: 2023-10-22.

[80] **Cofense**. Phishing Review. `https://go.cofense.com/wp-content/uploads/pdf/Cofense-Q3_2020_Phishing-Review-report.pdf`, 2020. Accessed: 2023-04-14.

[81] **Colosi, R.** Negatively Worded Questions Cause Respondent Confusion. *Proceedings of the Survey Research Methods Section* (2005), 2896–2903. Publisher: American Statistical Association.

[82] **Conway, D., Taib, R., Harris, M., Yu, K., Berkovsky, S., and Chen, F.** A Qualitative Investigation of Bank Employee Experiences of Information Security and Phishing. In *Proceedings of the 11th Symposium on Usable Privacy and Security, SOUPS 2017* (2017), USENIX Association, pp. 115–129.

[83] **Conzola, V. C., and Wogalter, M. S.** A Communication–Human Information Processing (C–HIP) approach to warning effectiveness in the workplace. *Journal of Risk Research 4*, 4 (Oct. 2001), 309–322. Publisher: Routledge.

[84] **Cranor, L. F.** A framework for reasoning about the human in the loop. In *Proceedings of the 1st Conference on Usability, Psychology, and Security* (USA, Apr. 2008), UPSEC'08, USENIX Association, pp. 1–15.

[85] **Darwish, A., Zarka, A. E., and Aloul, F.** Towards understanding phishing victims' profile. In *2012 International Conference on Computer Systems and Industrial Informatics* (2012), IEEE, pp. 1–5.

[86] **De Neys, W., and Glumicic, T.** Conflict monitoring in dual process theories of thinking. *Cognition 106*, 3 (2008), 1248–1299.

[87] **Dehaene, S., and Naccache, L.** Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition 79*, 1-2 (2001), 1–37.

[88] **Dhamija, R., Tygar, J., and Hearst, M.** Why Phishing Works. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2006), CHI '06, ACM, pp. 581–590.

[89] **Dijk, T.** Context and cognition. In *Discourse and Context: A Sociocognitive Approach*. Cambridge University Press, 2008, pp. 56–110.

[90] **Dijksterhuis, A., and Bargh, J.** The perception–behavior expressway: Automatic effects of social perception on social behavior. In *Advances in experimental social psychology*, vol. 33. Academic Press, 2001, pp. 1–40.

[91] **Dimoka, A., Pavlou, P. A., and Davis, F. D.** Research Commentary—NeuroIS: The Potential of Cognitive Neuroscience for Information Systems Research. *Information Systems Research 22*, 4 (Dec. 2011), 687–702. Publisher: INFORMS.

[92] **Ding, Y., Luktarhan, N., Li, K., and Slamu, W.** A keyword-based combination approach for detecting phishing webpages. *Computers & Security 84* (2019), 256–275.

[93] **Distler, V.** The Influence of Context on Response to Spear-Phishing Attacks: an In-Situ Deception Study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023), CHI '23, ACM, pp. 1–18.

[94] **DMOZ**. The Directory of the Web. `https://web.archive.org/web/20230326051429/https://dmoz-odp.org/`, 2017. Accessed: 2023-04-23.

[95] **Dobolyi, D. G., Abbasi, A., Zahedi, F. M., and Vance, A.** An Ordinal Approach to Modeling and Visualizing Phishing Susceptibility. In *International Conference on Intelligence and Security Informatics* (2020), IEEE, pp. 1–6.

[96] **Dodge, R., Carver, C., and Ferguson, A.** Phishing for user security awareness. *Computers & Security 26*, 1 (2007), 73–80.

[97] **Dodge, R., Coronges, K., and Rovira, E.** Empirical Benefits of Training to Phishing Susceptibility. In *Information Security and Privacy Research* (Berlin, Heidelberg, 2012), D. Gritzalis, S. Furnell, and M. Theoharidou, Eds., IFIP Advances in Information and Communication Technology, Springer, pp. 457–464.

[98] **Dolan, P., Hallsworth, M., Halpern, D., King, D., and Vlaev, I.** MINDSPACE Influencing behaviour through public policy. Tech. rep., Institute for Government, UK Cabinet Office, 2010.

[99] **Downs, J., Holbrook, M., and Cranor, L.** Decision strategies and susceptibility to phishing. In *Proceedings of the 2nd Symposium on Usable Privacy and Security, SOUPS 2009* (2006), ACM, pp. 79–90.

[100] **Dreibelbis, R. C.** It's More Than Just Changing Your Password: Exploring the Nature and Antecedents of Cyber-Security Behaviors. *USF Tampa Graduate Theses and Dissertations* (2016).

[101] **Duman, S., Kalkan-Cakmakci, K., Egele, M., Robertson, W., and Kirda, E.** EmailProfiler: Spearphishing Filtering with Header and Stylometric Features of Emails. In *Annual Computer Software and Applications Conference* (2016), vol. 1, pp. 408–416.

[102] **DutchReview**. You're under arrest: thousands of Dutchies targeted by phishing calls. `https://edu.nl/gwjpg`, 2021. Accessed: 2023-10-22.

[103] **Edwards, M., Larson, R., Green, B., Rashid, A., and Baron, A.** Panning for gold: Automatically analysing online social engineering attack surfaces. *Computers & Security 69* (2017), 18–34. Publisher: Elsevier.

[104] **Egelman, S., Cranor, L. F., and Hong, J.** You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2008), CHI '08, ACM, pp. 1065–1074.

[105] **Egelman, S., and Schechter, S.** The Importance of Being Earnest [In Security Warnings]. In *Financial Cryptography and Data Security* (Berlin, Heidelberg, 2013), A.-R. Sadeghi, Ed., LNCS, Springer, pp. 52–59.

[106] **Elsevier**. Scopus Content Coverage Guide. `https://edu.nl/wxmgh`, 2020. Accessed: 2023-10-22.

[107] **ENISA**. Threat Landscape for Supply Chain Attacks. Report/Study, European Agency for Cyber-Security, July 2021.

[108] **ENISA**. Threat Landscape. Report/Study, European Union Agency for Cyber Security, 2022.

[109] **European Commision**. Internal Market, Industry, Entrepreneurship and SMEs. `https://single-market-economy.ec.europa.eu/smes_en`. Accessed: 2023-03-20.

[110] **Evangelista, J. R. G., Sassi, R. J., Romero, M., and Napolitano, D.** Systematic Literature Review to Investigate the Application of Open Source Intelligence (OSINT) with Artificial Intelligence. *Journal of Applied Security Research 16*, 3 (July 2021), 345–369.

[111] **Evans, J.** In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences 7*, 10 (2003), 454–459.

[112] **F-Secure**. Attack Landscape Update 2021. `https://blog-assets.f-secure.com/wp-content/uploads/2021/03/30120359/attack-landscape-update-h1-2021.pdf`, 2021. Accessed: 2023-10-22.

[113] **Ferguson, A. J.** Fostering E-Mail Security Awareness: The West Point Carronade. *Educause Quarterly 28*, 1 (2005), 54–57.

[114] **Ferreira, A., Coventry, L., and Lenzini, G.** Principles of Persuasion in Social Engineering and Their Use in Phishing. In *Human Aspects of Information Security, Privacy, and Trust* (2015), LNCS, Springer, pp. 36–47.

[115] **Ferreira, A., and Lenzini, G.** An analysis of social engineering principles in effective phishing. In *2015 Workshop on Socio-Technical Aspects in Security and Trust* (2015), IEEE, pp. 9–16.

[116] **Ferreira, A. M., and Marques, P. M. V.** Phishing Through Time: A Ten Year Story based on Abstracts. In *ICISSP* (2018), pp. 225–232.

[117] **Fink, E., Sharifi, M., and Carbonell, J. G.** Application of machine learning and crowdsourcing to detection of cybersecurity threats. In *Proceedings of the US Department of Homeland Security Science Conference–Fifth Annual University Network Summit, Washington, DC* (2011).

[118] **Fisher, R. A.** Statistical Methods for Research Workers. In *Breakthroughs in Statistics: Methodology and Distribution*. Springer, New York, NY, 1992, pp. 66–70.

[119] **Flores, W., Holm, H., Nohlberg, M., and Ekstedt, M.** Investigating personal determinants of phishing and the effect of national culture. *Information & Computer Security 23*, 2 (Jan. 2015), 178–199. Publisher: Emerald Group Publishing Limited.

[120] **Flores, W., Holm, H., Svensson, G., and Ericsson, G.** Using phishing experiments and scenario-based surveys to understand security behaviours in practice. In *European Information Security Multi-Conference* (2013), AISEL, pp. 79–90.

[121] **Franz, A., Zimmermann, V., Albrecht, G., Hartwig, K., Reuter, C., Benlian, A. r., and Vogt, J.** SoK: Still Plenty of Phish in the Sea — A Taxonomy of User-Oriented Phishing Interventions and Avenues for Future Research. In *Proceedings of the 17th Symposium on Usable Privacy and Security, SOUPS 2021* (2021), USENIX Association, pp. 339–358.

[122] **Fu, A. Y., Wenyin, L., and Deng, X.** Detecting Phishing Web Pages with Visual Similarity Assessment Based on Earth Mover's Distance (EMD). *IEEE Transactions on Dependable and Secure Computing 3*, 4 (Oct. 2006), 301–311. Conference Name: IEEE Transactions on Dependable and Secure Computing.

[123] **Gershman, A., McCarthy, J., and Fano, A.** Situated computing: bridging the gap between intention and action. In *International Symposium on Wearable Computers* (1999), IEEE, pp. 3–9.

[124] **Gigerenzer, G.** How to Make Cognitive Illusions Disappear: Beyond "Heuristics and Biases". *European Review of Social Psychology 2*, 1 (1991), 83–115.

[125] **Goel, S., Williams, K., and Dincelli, E.** Got phished? Internet security and human vulnerability. *Journal of the Association for Information Systems 18*, 1 (2017), 22–44.

[126] **Goldberg, L. R.** A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe 7* (1999), 7–28.

[127] **Google**. Understanding why phishing attacks are so effective and how to mitigate them. `https://security.googleblog.com/2019/08/understanding-why-phishing-attacks-are.html`, 2019. Accessed: 2019-08-08.

[128] **Google**. Google Safe Browsing. `https://safebrowsing.google.com/`, 2021. Accessed: 2020-11-16.

[129] **Grazioli, S.** Where Did They Go Wrong? An Analysis of the Failure of Knowledgeable Internet Consumers to Detect Deception Over the Internet. *Group Decision and Negotiation 13*, 2 (2004), 149–172.

[130] **Grazioli, S., and Jarvenpaa, S.** Perils of Internet fraud: an empirical investigation of deception and trust with experienced Internet consumers. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 30*, 4 (July 2000), 395–410. Conference Name: IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans.

[131] **Grazioli, S., and Wang, A.** Looking Without Seeing: Understanding Unsophisticated Consumers' Success and Failure to Detect Internet Deception. In *ICIS* (2001), AISEL, p. 13.

[132] **Greene, K., Steves, M. P., Theofanos, M. F., and Kostick, J. A.** User Context: An Explanatory Variable in Phishing Susceptibility. In *Network and Distributed Systems Security (NDSS) Symposium* (July 2018), Internet Society.

[133] **Greening, T.** Ask and ye shall receive: a study in social engineering. *ACM SIGSAC Review 14*, 2 (1996), 8–14.

[134] **Griggs, R. A., and Cox, J. R.** The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology 73*, 3 (1982), 407–420.

[135] **Guo, K. H.** Security-related behavior in using information systems in the workplace: A review and synthesis. *Computers & Security 32* (2013), 242–251.

[136] **Gupta, B. B., Arachchilage, N. A. G., and Psannis, K. E.** Defending against phishing attacks: taxonomy of methods, current issues and future directions. *Telecommunication Systems 67*, 2 (Feb. 2018), 247–267.

[137] **Gupta, S., Gupta, P., Ahamad, M., and Kumaraguru, P.** Exploiting Phone Numbers and Cross-Application Features in Targeted Mobile Attacks. In *Workshop on Security and Privacy in Smartphones and Mobile Devices* (New York, NY, USA, 2016), ACM, pp. 73–82.

[138] **Gupta, S., and Kumaraguru, P.** Emerging phishing trends and effectiveness of the anti-phishing landing page. In *2014 APWG Symposium on Electronic Crime Research (eCrime)* (Sept. 2014), IEEE, pp. 36–47. ISSN: 2159-1245.

[139] **Hae, K. J.** Multicollinearity and misleading statistical results. *Korean J Anesthesiol 72*, 6 (2019), 558–569.

[140] **Hale, M., Gamble, R., and Gamble, P.** CyberPhishing: A Game-Based Platform for Phishing Awareness Testing. In *Hawaii International Conference on System Sciences* (2015), IEEE, pp. 5260–5269.

[141] **Halevi, T., Memon, N., Lewis, J., Kumaraguru, P., Arora, S., Dagar, N., Aloul, F., and Chen, J.** Cultural and Psychological Factors in Cyber-Security. In *International Conference on Information Integration and Web-based Applications & Services* (New York, NY, USA, 2016), ACM, pp. 43–56.

[142] **Halevi, T., Memon, N., and Nov, O.** Spear-Phishing in the Wild: A Real-World Study of Personality, Phishing Self-Efficacy and Vulnerability to Spear-Phishing Attacks. *SSRN Electronic Journal ssrn.2544742* (2015).

[143] **Han, X., Kheir, N., and Balzarotti, D.** PhishEye: Live Monitoring of Sandboxed Phishing Kits. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (New York, NY, USA, Oct. 2016), CCS '16, ACM, pp. 1402–1413.

[144] **Hardy, S., Crete-Nishihata, M., Kleemola, K., Senft, A., Sonne, B., Wiseman, G., Gill, P., and Deibert, R. J.** Targeted Threat Index: Characterizing and Quantifying Politically-Motivated Targeted Malware. In *USENIX Security Symposium* (2014), USENIX Association, pp. 527–541.

[145] **Harrison, B., Svetieva, E., and Vishwanath, A.** Individual processing of phishing emails: How attention and elaboration protect against phishing. *Online Information Review 40*, 2 (2016), 265–281.

[146] **Hastie, R., and Dawes, R. M.** *Rational Choice in an Uncertain World: The Psychology of Judgment and Decision Making*. SAGE, 2010.

[147] **Heartfield, R., and Loukas, G.** A Taxonomy of Attacks and a Survey of Defence Mechanisms for Semantic Social Engineering Attacks. *ACM Computing Surveys 48*, 3 (2016), 1–39.

[148] **Ho, G., Cidon, A., Gavish, L., Schweighauser, M., Paxson, V., Savage, S., Voelker, G., and Wagner, D.** Detecting and Characterizing Lateral Phishing at Scale. In *USENIX Security Symposium* (2019), USENIX Association, pp. 1273–1290.

[149] **Ho, G., Sharma, A., Javed, M., Paxson, V., and Wagner, D.** Detecting Credential Spearphishing in Enterprise Settings. In *USENIX Security Symposium* (2017), USENIX Association, pp. 469–485.

[150] **Holm, H., Flores, W., Nohlberg, M., and Ekstedt, M.** An empirical investigation of the effect of target-related information in phishing attacks. In *International Enterprise Distributed Object Computing Workshop* (2014), IEEE, pp. 357–363.

[151] **Holm, H., Flores, W. R., and Ericsson, G.** Cyber security for a Smart Grid - What about phishing? In *IEEE PES ISGT Europe* (2013), IEEE, pp. 1–5.

[152] **Hong, J.** The state of phishing attacks. *Communications of the ACM 55*, 1 (Jan. 2012), 74–81.

[153] **Houdé, O., Zago, L., Mellet, E., Moutier, S., Pineau, A., Mazoyer, B., and Tzourio-Mazoyer, N.** Shifting from the Perceptual Brain to the Logical Brain: The Neural Impact of Cognitive Inhibition Training. *Journal of Cognitive Neuroscience 12*, 5 (Sept. 2000), 721–728. Conference Name: Journal of Cognitive Neuroscience.

[154] **House, D., and Raja, M. K.** Phishing: message appraisal and the exploration of fear and self-confidence. *Behaviour and Information Technology 39*, 11 (Nov. 2020), 1204–1224. Publisher: Taylor & Francis.

[155] **Hu, H., and Wang, G.** Revisiting Email Spoofing Attacks. *CoRR arXiv:1801.00853* (2018).

[156] **Hutcheson, G. D.** Ordinary least-squares regression. In *The SAGE dictionary of quantitative management research*. SAGE Knowledge, 2011, pp. 224–228.

[157] **IBM Corp.** IBM SPSS Statistics for Windows. `https://www.ibm.com/products/spss-statistics`, 2021. Accessed: 2022-5-22.

[158] **Irani, D., Balduzzi, M., Balzarotti, D., Kirda, E., and Pu, C.** Reverse Social Engineering Attacks in Online Social Networks. In *Detection of Intrusions and Malware, and Vulnerability Assessment* (2011), LNCS, Springer, pp. 55–74.

[159] **Iuga, C., Nurse, J. R. C., and Erola, A.** Baiting the hook: factors impacting susceptibility to phishing attacks. *Human-centric Computing and Information Sciences 6*, 1 (June 2016), 8.

[160] **Jackson, C., Simon, D., Tan, D. S., and Barth, A.** An Evaluation of Extended Validation and Picture-in-Picture Phishing Attacks. In *Financial Cryptography and Data Security* (2007), LNCS, Springer, pp. 281–293.

[161] **Jagatic, T. N., Johnson, N. A., Jakobsson, M., and Menczer, F.** Social phishing. *Communications of the ACM 50*, 10 (Oct. 2007), 94–100. Publisher: ACM.

[162] **Jahangir, N., Akbar, M. M., and Haq, M.** Organisational Citizenship Behavior: Its Nature and Antecedents. *BRAC University Journal 1*, 2 (2004), 75–85.

[163] **Jain, A., and Gupta, B.** Phishing Detection: Analysis of Visual Similarity Based Approaches. *Security and Communication Networks 2017* (Jan. 2017), 1–20.

[164] **Jakobsson, M., and Ratkiewicz, J.** Designing ethical phishing experiments: a study of (ROT13) rOnl query features. In *Proceedings of the 15th international conference on World Wide Web* (New York, NY, USA, May 2006), WWW '06, ACM, pp. 513–522.

[165] **Jakobsson, M., Tsow, A., Shah, A., Blevis, E., and Lim, Y.-K.** What Instills Trust? A Qualitative Study of Phishing. In *Financial Cryptography and Data Security* (Berlin, Heidelberg, 2007), S. Dietrich and R. Dhamija, Eds., vol. 4886 of *LNCS*, Springer, pp. 356–361.

[166] **Jalali, M. S., Bruckes, M., Westmattelmann, D., and Schewe, G.** Why Employees (Still) Click on Phishing Links: Investigation in Hospitals. *Journal of Medical Internet Research 22*, 1 (Jan. 2020), e16775. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[167] **Jampen, D., Gür, G., Sutter, T., and Tellenbach, B.** Don't click: towards an effective anti-phishing training. A comparative literature review. *Human-centric Computing and Information Sciences 10*, 1 (Aug. 2020), 33.

[168] **Janke, R.** Effects of Mentioning the Incentive Prize in the Email Subject Line on Survey Response. *Evidence Based Library and Information Practice 9*, 1 (Mar. 2014), 4–13. Number: 1.

[169] **Jansson, K., and Solms, R. v.** Phishing for phishing awareness. *Behaviour & Information Technology 32*, 6 (June 2013), 584–593. Publisher: Taylor & Francis.

[170] **Jensen, M. L., Dinger, M., Wright, R. T., and Thatcher, J. B.** Training to Mitigate Phishing Attacks Using Mindfulness Techniques. *Journal of Management Information Systems 34*, 2 (Apr. 2017), 597–626. Publisher: Routledge.

[171] **Jensen, M. L., Wright, R. T., Durcikova, A., and Karumbaiah, S.** Improving Phishing Reporting Using Security Gamification. *Journal of Management Information Systems 39*, 3 (2022), 793–823. Publisher: Routledge.

[172] **John, O. P., Donahue, E. M., and Kentle, R. L.** Big five inventory. *Journal of Personality and Social Psychology 1*, 1 (1991).

[173] **Jones, N. A., Ross, H., Lynam, T., Perez, P., and Leitch, A.** Mental Models: An Interdisciplinary Synthesis of Theory and Methods. *Ecology and Society 16*, 1 (2011).

[174] **Junger, M., Montoya, L., and Overink, F.-J.** Priming and warnings are not effective to prevent social engineering attacks. *Computers in Human Behavior 66* (2017), 75–87.

[175] **Kahneman, D.** A perspective on judgment and choice: Mapping bounded rationality. *American psychologist* (2003), 697–720.

[176] **Kahneman, D.** *Thinking, Fast and Slow*. Farrar, Straus and Giroux, Oct. 2011.

[177] **Kearney, W., and Kruger, H.** Can perceptual differences account for enigmatic information security behaviour in an organisation? *Computers & Security 61* (2016), 46–58.

[178] **Kent, R., and Brandal, H.** Improving email response in a permission marketing context. *International Journal of Market Research 45*, 4 (2003), 1–13.

[179] **Kersten, L., Burda, P., Allodi, L., and Zannone, N.** Investigating the Effect of Phishing Believability on Phishing Reporting. In *2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)* (June 2022), IEEE, pp. 117–128.

[180] **Khonji, M., Iraqi, Y., and Jones, A.** Mitigation of spear phishing attacks: A Content-based Authorship Identification framework. In *International Conference for Internet Technology and Secured Transactions* (2011), pp. 416–421.

[181] **Khonji, M., Iraqi, Y., and Jones, A.** Phishing Detection: A Literature Survey. *IEEE Communications Surveys & Tutorials 15*, 4 (2013), 2091–2121.

[182] **Kim, B., Lee, D.-Y., and Kim, B.** Deterrent effects of punishment and training on insider security threats: a field experiment on phishing attacks. *Behaviour and Information Technology 39*, 11 (Nov. 2020), 1156–1175. Publisher: Taylor & Francis.

[183] **Kindberg, T., O'Neill, E., Bevan, C., Kostakos, V., Fraser, D., and Jay, T.** Measuring Trust in Wi-Fi hotspots. In *Conference on Human Factors in Computing Systems* (2008), ACM, pp. 173–182.

[184] **Kleitman, S., Law, M., and Kay, J.** It's the deceiver and the receiver: Individual differences in phishing susceptibility and false positives with item profiling. *PLOS ONE 13*, 10 (2018), 1–29.

[185] **Kline, P.** *The Handbook of Psychological Testing (2nd ed.)*. Routledge, 1993.

[186] **Knez, I., Hjärpe, D., and Bryngelsson, M.** Predicting Organizational Citizenship Behavior: The Role of Work-Related Self. *SAGE Open 9* (2019), 215824401985483.

[187] **KnowBe4**. Phishing. `https://www.knowbe4.com/phishing`, 2021. Accessed: 2023-10-22.

[188] **Kokulu, F. B., Soneji, A., Bao, T., Shoshitaishvili, Y., Zhao, Z., Doupé, A., and Ahn, G.-J.** Matched and Mismatched SOCs: A Qualitative Study on Security Operations Center Issues. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* (New York, NY, USA, 2019), CCS '19, ACM, pp. 1955–1970.

[189] **Kotov, V., and Massacci, F.** Anatomy of Exploit Kits. In *Engineering Secure Software and Systems* (2013), LNC&S, Springer, pp. 181–196.

[190] **Krizhevsky, A., Sutskever, I., and Hinton, G.** ImageNet Classification with Deep Convolutional Neural Networks. *Comm. ACM* (2017), 84–90. Publisher: ACM.

[191] **Krosnick, J. A., and Presser, S.** Question and questionnaire design. In *Handbook of survey research*. Emerald, 2010, pp. 263–314.

[192] **Kumaraguru, P., Cranshaw, J., Acquisti, A., Cranor, L., Hong, J., Blair, M., and Pham, T.** School of phish: A real-world evaluation of anti-phishing training. In *Proceedings of the 5th Symposium on Usable Privacy and Security, SOUPS 2009* (2009), ACM, pp. 1–12.

[193] **Kumaraguru, P., Rhee, Y., Sheng, S., Hasan, S., Acquisti, A., Cranor, L., and Hong, J.** Getting users to pay attention to anti-phishing education: Evaluation of retention and transfer. In *International Conference Proceeding Series* (2007), vol. 269, ACM, pp. 70–81.

[194] **Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L., and Hong, J.** Lessons from a real world evaluation of anti-phishing training. In *Symposium on Electronic Crime Research* (2008), IEEE, pp. 1–12.

[195] **Kwak, Y., Lee, S., Damiano, A., and Vishwanath, A.** Why do users not report spear phishing emails? *Telematics and Informatics 48* (May 2020), 101343.

[196] **Lain, D., Kostiainen, K., and Čapkun, S.** Phishing in Organizations: Findings from a Large-Scale and Long-Term Study. In *Symposium on Security and Privacy (SP)* (2022), IEEE, pp. 842–859.

[197] **Langner, R.** Stuxnet: Dissecting a Cyberwarfare Weapon. *IEEE Security and Privacy 9*, 3 (May 2011), 49–51. Conference Name: IEEE Security Privacy.

[198] **Langner, R., and Group, T. L.** To Kill a Centrifuge - A Technical Analysis of What Stuxnet's Creators Tried to Achieve. Tech. rep., The Langner Group, 2013.

[199] **Lastdrager, E., Gallardo, I., Hartel, P., and Junger, M.** How effective is anti-phishing training for children? In *Proceedings of the 13th Symposium on Usable Privacy and Security, SOUPS 2017* (2019), USENIX Association, pp. 229–239.

[200] **Lastdrager, E. E.** Achieving a consensual definition of phishing based on a systematic review of the literature. *Crime Science 3*, 1 (Sept. 2014), 9.

[201] **Lavie, N.** Distracted and confused?: Selective attention under load. *Trends in Cognitive Sciences 9*, 2 (2005), 75–82.

[202] **Lelkes, Y., and Weiss, R.** Much ado about acquiescence: The relative validity and reliability of construct-specific and agree–disagree questions. *Research & Politics 2*, 3 (2015).

[203] **LePine, J., Erez, A., and Johnson, D.** The Nature and Dimensionality of Organizational Citizenship Behavior: A Critical Review and Meta-Analysis. *The Journal of applied psychology 87* (2002), 52–65.

[204] **Leusden, F. v.** Why our awareness campaigns are bad, and we should feel bad. https://one-conference.nl/video/wf-19-why-our-awareness-campaigns-are-bad-and-we-should-feel-bad.mp4, 2022. Accessed: 2023-03-08; Publisher: ONE Conference 2022.

[205] **Li, F., Durumeric, Z., Czyz, J., Karami, M., Bailey, M., McCoy, D., Savage, S., and Paxson, V.** You've got vulnerability: Exploring effective vulnerability notifications. In *USENIX Security Symposium* (2016), USENIX Association, pp. 1033–1050.

[206] **Li, Y., Xiong, K., and Li, X.** An analysis of user behaviors in phishing email using machine learning techniques. In *International Joint Conference on e-Business and Telecommunications* (2019), vol. 2, SCITEPRESS, pp. 529–534.

[207] **Lien, M., Allen, P., Ruthruff, E., Grabbe, J., McCann, R., and Remington, R.** Visual word recognition without central attention: Evidence for greater automaticity with advancing age. *Psychology and Aging 21*, 3 (2006), 431–447.

[208] **Lin, E., Greenberg, S., Trotter, E., Ma, D., and Aycock, J.** Does domain highlighting help people identify phishing sites? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2011), CHI '11, ACM, pp. 2075–2084.

[209] **Lin, T., Capecci, D. E., Ellis, D. M., Rocha, H. A., Dommaraju, S., Oliveira, D. S., and Ebner, N. C.** Susceptibility to Spear-Phishing Emails: Effects of Internet User Demographics and Email Content. *ACM Transactions on Computer-Human Interaction 26*, 5 (July 2019), 32:1–32:28.

[210] **Lin, Y., Liu, R., Divakaran, D. M., Ng, J. Y., Chan, Q. Z., Lu, Y., Si, Y., Zhang, F., and Dong, J. S.** Phishpedia: A Hybrid Deep Learning Based Approach to Visually Identify Phishing Webpages. In *USENIX Security Symposium* (2021), USENIX Association, pp. 3793–3810.

[211] **Liu, G., Qiu, B., and Liu, W.** Automatic Detection of Phishing Target from Phishing Webpage. In *International Conference on Pattern Recognition* (2010), pp. 4153–4156.

[212] **Liu, G., Xiang, G., Pendleton, B. A., Hong, J. I., and Liu, W.** Smartening the crowds: computational techniques for improving human verification to fight phishing scams. In *Proceedings of the 7th Symposium on Usable Privacy and Security, SOUPS 2011* (2011), ACM.

[213] **Liu, R., Lin, Y., Yang, X., Ng, S. H., Divakaran, D. M., and Dong, J. S.** Inferring Phishing Intention via Webpage Appearance and Dynamics: A Deep Vision Based Approach. In *USENIX Security Symposium* (2022), USENIX Association, pp. 1633–1650.

[214] **London City University**. 10 Rules for Writing Professional Emails. https://www.city.ac.uk/__data/assets/pdf_file/0003/234354/Writing-Professional-Emails.pdf, Oct. 2015. Accessed: 2019-03-19.

[215] **Luo, X. R., Zhang, W., Burd, S., and Seazzu, A.** Investigating phishing victimization with the Heuristic–Systematic Model: A theoretical framework and an exploration. *Computers & Security 38* (Oct. 2013), 28–38.

[216] **Ma, Z., Reynolds, J., Dickinson, J., Wang, K., Judd, T., Barnes, J., Mason, J., and Bailey, M.** The impact of secure transport protocols on phishing efficacy. In *USENIX Workshop on Cyber Security Experimentation and Test* (2019), USENIX Association.

[217] **MacLean, K. A., Aichele, S. R., Bridwell, D. A., Mangun, G. R., Wojciulik, E., and Saron, C. D.** Interactions between Endogenous and Exogenous Attention during Vigilance. *Attention, perception & psychophysics 71*, 5 (2009), 1042–1058.

[218] **MacPherson, K.** *Permission-based E-mail Marketing that Works!* Dearborn Trade Publishing, 2001.

[219] **Marchal, S., Saari, K., Singh, N., and Asokan, N.** Know Your Phish: Novel Techniques for Detecting Phishing Sites and their Targets. *International Conference on Distributed Computing Systems* (2015), 323–333.

[220] **Marczak, W. R., Scott-Railton, J., Marquis-Boire, M., and Paxson, V.** When Governments Hack Opponents: A Look at Actors and Technology. In *USENIX Security Symposium* (2014), USENIX Association, pp. 511–525.

[221] **Marett, K., and Wright, R.** The effectiveness of deceptive tactics in phishing. In *Americas Conference on Information Systems* (2009), vol. 4, AISEL, pp. 2583–2591.

[222] **Marin, I. A., Burda, P., Zannone, N., and Allodi, L.** The Influence of Human Factors on the Intention to Report Phishing Emails. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, Apr. 2023), CHI '23, ACM, pp. 1–18.

[223] **Martin, J., Dubé, C., and Coovert, M. D.** Signal Detection Theory (SDT) Is Effective for Modeling User Behavior Toward Phishing and Spear-Phishing Attacks. *Human Factors: The Journal of the Human Factors and Ergonomics Society 60*, 8 (2018), 1179–1191.

[224] **Martínez-Mesa, J., González-Chica, D., Bastos, J., Bonamigo, R., and Duquia, R.** Sample size: how many participants do I need in my research? *Anais brasileiros de dermatologia 89* (2014), 609–615.

[225] **Mayer, P., Kunz, A., and Volkamer, M.** Reliable Behavioural Factors in the Information Security Context. In *International Conference on Availability, Reliability and Security* (New York, NY, USA, 2017), ACM.

[226] **McBee, E., Ratcliffe, T., Picho, K., Artino, A. R., Schuwirth, L., Kelly, W., Masel, J., van der Vleuten, C., and Durning, S. J.** Consequences of contextual factors on clinical reasoning in resident physicians. *Advances in Health Sciences Education 20*, 5 (Dec. 2015), 1225–1236.

[227] **McCrae, R. R., and Costa, P. T.** Validation of the Five-Factor Model of Personality Across Instruments and Observers. *Journal of Personality and Social Psychology 52*, 1 (1987), 81–90. Publisher: American Psychological Association Inc.

[228] **McDonald, N., Schoenebeck, S., and Forte, A.** Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction 3*, CSCW (Nov. 2019), 72:1–72:23.

[229] **Medvet, E., Kirda, E., and Kruegel, C.** Visual-Similarity-Based Phishing Detection. In *International Conference on Security and Privacy in Communication Networks* (2008), ACM.

[230] **Merriam, S.** *Qualitative research and case study applications in education*. Jossey-Bass Publishers, 1998.

[231] **Michie, S., West, R., Campbell, R., Brown, J., and Gainforth, H.** *ABC of Behaviour Change Theories*. Silverback Publishing, 2014.

[232] **Microsoft**. Microsoft Digital Defense Report. `https://www.microsoft.com/en-us/security/business/microsoft-digital-defense-report-2021`, 2021. Accessed: 2021-10-01.

[233] **Mitnick, K., and Simon, W.** *The Art of Deception: Controlling the Human Element of Security*. John Wiley & Sons, 2003.

[234] **Miyamoto, D., Iimura, T., Blanc, G., Tazaki, H., and Kadobayashi, Y.** EyeBit: Eye-Tracking Approach for Enforcing Phishing Prevention Habits. In *International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security* (2016), IEEE, pp. 56–65.

[235] **Modic, D., and Anderson, R.** Reading this may harm your computer: The psychology of malware warnings. *Computers in Human Behavior 41* (2014), 71–79.

[236] **Mohammad, R., and McCluskey, L.** Phishing Websites Data Set. `https://archive.ics.uci.edu/ml/datasets/phishing+websites/`, 2015. Accessed: 2023-10-22; Publisher: University of California.

[237] **Molinaro, K., and Bolton, M.** Evaluating the applicability of the double system lens model to the analysis of phishing email judgments. *Computers & Security 77* (2018), 128–137.

[238] **Montañez, R., Golob, E., and Xu, S.** Human Cognition Through the Lens of Social Engineering Cyberattacks. *Frontiers in Psychology 11* (2020). Publisher: Frontiers.

[239] **Moody, G., Galletta, D., and Dunn, B.** Which phish get caught An exploratory study of individuals' susceptibility to phishing. *European Journal of Information Systems 26*, 6 (2017), 564–584.

[240] **Moore, T., and Anderson, R.** How brain type influences online safety. In *Workshop on Security and Human Behaviour* (2008), University of Cambridge, p. 8.

[241] **Moore, T., and Clayton, R.** Evaluating the Wisdom of Crowds in Assessing Phishing Websites. In *Financial Cryptography and Data Security* (2008), LNCS, Springer, pp. 16–30.

[242] **Morgan, P., Williams, E., Zook, N., and Christopher, G.** Exploring Older Adult Susceptibility to Fraudulent Computer Pop-Up Interruptions. In *Advances in Intelligent Systems and Computing* (2019), vol. 782, Springer, pp. 56–68.

[243] **Muppavarapu, V., Rajendran, A., and Vasudevan, S.** Phishing detection using RDF and random forests. *Int. Arab. J. Inf. Technol. 15* (2018), 817–824.

[244] **Musuva, P., Getao, K., and Chepken, C.** A new approach to modelling the effects of cognitive processing and threat detection on phishing susceptibility. *Computers in Human Behavior 94* (2019), 154–175.

[245] **Muthal, S., Li, S., Huang, Y., Li, X., Dahbura, A., Bos, N., and Molinaro, K.** A phishing study of user behavior with incentive and informed intervention. In *Annual Cyber Security Summit* (2017), Digital Commons.

[246] **Ndibwile, J. D., Luhanga, E. T., Fall, D., Miyamoto, D., Blanc, G., and Kadobayashi, Y.** An Empirical Approach to Phishing Countermeasures Through Smart Glasses and Validation Agents. *IEEE Access 7* (2019), 130758–130771. Conference Name: IEEE Access.

[247] **Neupane, A., Saxena, N., and Hirshfield, L.** Neural Underpinnings of Website Legitimacy and Familiarity Detection: An fNIRS Study. In *Proceedings of the 26th International Conference on World Wide Web* (Republic and Canton of Geneva, CHE, Apr. 2017), WWW '17, International World Wide Web Conferences Steering Committee, pp. 1571–1580.

[248] **Neupane, A., Saxena, N., Maximo, J. O., and Kana, R.** Neural Markers of Cybersecurity: An fMRI Study of Phishing and Malware Warnings. *IEEE Transactions on Information Forensics and Security 11*, 9 (Sept. 2016), 1970–1983. Conference Name: IEEE Transactions on Information Forensics and Security.

[249] **Nicholson, J., Coventry, L., and Briggs, P.** Can We Fight Social Engineering Attacks By Social Means? Assessing Social Salience as a Means to Improve Phish Detection. In *Proceedings of the 11th Symposium on Usable Privacy and Security, SOUPS 2017* (2017), USENIX Association, p. 15.

[250] **Nicholson, J., Javed, Y., Dixon, M., Coventry, L., Ajayi, O. D., and Anderson, P.** Investigating Teenagers' Ability to Detect Phishing Messages. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS PW)* (Sept. 2020), IEEE, pp. 140–149.

[251] **Nielsen, T. M., Bachrach, D. G., Sundstrom, E., and Halfhill, T. R.** Utility of OCB: Organizational Citizenship Behavior and Group Performance in a Resource Allocation Framework. *Journal of Management 38*, 2 (2012), 668–694.

[252] **NIST**. The Five Functions. Tech. rep., National Institute of Standards and Technology, Apr. 2018.

[253] **Norman, W. T.** Toward an adequate taxonomy of personality attributes: replicated factors structure in peer nomination personality ratings. *Journal of abnormal and social psychology 66* (1963), 574–583.

[254] **Nosek, B., Hawkins, C., and Frazier, R.** Implicit social cognition: From measures to mechanisms. *Trends in Cognitive Sciences 15*, 4 (2011), 152–159.

[255] **Oest, A., Zhang, P., Wardman, B., Nunes, E., Burgis, J., Zand, A., Thomas, K., Doupé, A., and Ahn, G.-J.** Analyzing the End-to-end Life Cycle and E ectiveness of Phishing Attacks at Scale. In *USENIX Security Symposium* (2021), USENIX Association, p. 17.

[256] **Oliveira, D., Rocha, H., Yang, H., Ellis, D., Dommaraju, S., Muradoglu, M., Weir, D., Soliman, A., Lin, T., and Ebner, N.** Dissecting Spear Phishing Emails for Older vs Young Adults: On the Interplay of Weapons of Influence and Life Domains in Predicting Susceptibility to Phishing. In *Conference on Human Factors in Computing Systems* (2017), ACM, pp. 6412–6424.

[257] **Onarlioglu, K., Yilmaz, U. O., Kirda, E., and Balzarotti, D.** Insights into User Behavior in Dealing with Internet Attacks. In *Network and Distributed Systems Security (NDSS) Smposium* (2012), Internet Society.

[258] **OpenPhish**. OpenPhish - Phishing Intelligence. `https://openphish.com/`, 2020. Accessed: 2020-04-09.

[259] **Oppenheimer, D.** The secret life of fluency. *Trends in Cognitive Sciences 12*, 6 (2008), 237–241.

[260] **Organ, D.** Organizational Citizenship Behavior: It's Construct Clean-Up Time. *Human Performance 10* (1997), 85–97.

[261] **Organ, D. W.** *Organizational citizenship behavior: The good soldier syndrome*. Lexington Books/D. C. Heath and Com, Lexington, MA, England, 1988. Pages: xiii, 132.

[262] **Otsu, N.** A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics 9*, 1 (1979), 62–66.

[263] **Pan, Y.and Xuhua, D.** Anomaly Based Web Phishing Page Detection. In *Annual Computer Security Applications Conference* (2006), pp. 381–392.

[264] **Parkin, S., Fielder, A., and Ashby, A.** Pragmatic Security: Modelling IT Security Management Responsibilities for SME Archetypes. In *International Workshop on Managing Insider Security Threats* (2016), ACM, pp. 69–80.

[265] **Parsons, K., Butavicius, M., Delfabbro, P., and Lillie, M.** Predicting susceptibility to social influence in phishing emails. *International Journal of Human-Computer Studies 128* (Aug. 2019), 17–26.

[266] **Parsons, K., McCormac, A., Pattinson, M., Butavicius, M., and Jerram, C.** Phishing for the Truth: A Scenario-Based Experiment of Users' Behavioural Response to Emails. In *Security and Privacy Protection in Information Processing Systems* (2013), vol. 405 of *IFIP Advances in Information and Communication Technology*, Springer, pp. 366–378.

[267] **Parsons, K., Mccormac, A., Pattinson, M., Butavicius, M., and Jerram, C.** Using Actions and Intentions to Evaluate Categorical Responses to Phishing and Genuine Emails. In *8th International Symposium on Human Aspects of Information Security & Assurance HAISA (2014)* (2014), Centre for Security, Communications & Network Research, Plymouth University.

[268] **Parsons, K., McCormac, A., Pattinson, M., Butavicius, M., and Jerram, C.** The design of phishing studies: Challenges for researchers. *Computers & Security 52* (2015), 194–206.

[269] **Patsakis, C., and Chrysanthou, A.** Analysing the fall 2020 Emotet campaign. *CoRR arXiv:2011.06479* (2020).

[270] **Pattinson, M., Jerram, C., Parsons, K., McCormac, A., and Butavicius, M.** Managing phishing emails: A scenario-based experiment. In *Proceedings of the 5th International Symposium on Human Aspects of Information Security and Assurance, HAISA 2011* (2011), Centre for Security, Communications & Network Research, Plymouth University, pp. 75–85.

[271] **Pattinson, M., Jerram, C., Parsons, K., McCormac, A., and Butavicius, M.** Why do some people manage phishing e-mails better than others? *Information Management & Computer Security 20*, 1 (2012), 18–28.

[272] **Peng, P., Xu, C., Quinn, L., Hu, H., Viswanath, B., and Wang, G.** What Happens After You Leak Your Password: Understanding Credential Sharing on Phishing Sites. In *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security* (New York, NY, USA, July 2019), Asia CCS '19, ACM, pp. 181–192.

[273] **Petelka, J., Zou, Y., and Schaub, F.** Put Your Warning Where Your Link Is: Improving and Evaluating Email Phishing Warnings. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), CHI '19, ACM, pp. 1–15.

[274] **Petrič, G., and Roer, K.** The impact of formal and informal organizational norms on susceptibility to phishing: Combining survey and field experiment data. *Telematics and Informatics* (2021), 101766. Publisher: Elsevier.

[275] **Pfeiffer, T., Kauer, M., and Röth, J.** *A bank would never write that! A qualitative study on E-mail trust decisions*. Gesellschaft für Informatik e.V., 2014. Accepted: 2017-07-26T10:59:33Z ISSN: 1617-5468.

[276] **Pfleeger, S., and Caputo, D.** Leveraging behavioral science to mitigate cyber security risk. *Computers & Security 31*, 4 (2012), 597–611.

[277] **Phishlabs**. Phishing Trends and Intelligence Report. https://edu.nl/evgrp, 2019. Accessed: 2023-09-14; Copyright: Fortra (PhishLabs) 2023;.

[278] **PhishStats**. PhishStats. https://phishstats.info/, 2021. Accessed: 2020-04-09.

[279] **PhishTank**. Join the fight against phishing. https://www.phishtank.com/, 2020. Accessed: 2023-10-22.

[280] **Pinker, S.** *How the Mind Works*, reissue ed. W. W. Norton & Company, 2009.

[281] **Pirocca, S., Allodi, L., and Zannone, N.** A Toolkit for Security Awareness Training Against Targeted Phishing. In *International Conference on Information Systems Security* (2020), Springer, pp. 137–159.

[282] **Pisarchik, A. N., Maksimenko, V. A., and Hramov, A. E.** From Novel Technology to Novel Applications: Comment on "An Integrated Brain-Machine Interface Platform With Thousands of Channels" by Elon Musk and Neuralink. *Journal of Medical Internet Research 21*, 10 (Oct. 2019), e16356. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[283] **Podsakoff, P. M., MacKenzie, S. B., Moorman, R. H., and Fetter, R.** Transformational leader behaviors and their effects on followers' trust in leader, satisfaction, and organizational citizenship behaviors. *The Leadership Quarterly 1*, 2 (1990), 107–142.

[284] **Pond III, S., Nacoste, R., Mohr, M., and Rodriguez, C.** The Measurement of Organizational Citizenship Behavior: Are We Assuming Too Much? *Journal of Applied Social Psychology 27* (2006), 1527–1544.

[285] **Posey, C., Roberts, T. L., and Lowry, P. B.** The Impact of Organizational Commitment on Insiders' Motivation to Protect Organizational Information Assets. *Journal of Management Information Systems 32*, 4 (2015), 179–214. Publisher: Taylor & Francis.

[286] **Purkait, S.** Phishing counter measures and their effectiveness - Literature review. *Information Management & Computer Security 20*, 5 (2012), 382–420.

[287] **Purkait, S., Kumar De, S., and Suar, D.** An empirical investigation of the factors that influence Internet user's ability to correctly identify a phishing website. *Information Management & Computer Security 22*, 3 (Jan. 2014), 194–234.

[288] **Ramesh, G., Krishnamurthi, I., and Kumar, K.** An efficacious method for detecting phishing webpages through target domain identification. *Decision Support Systems 61* (2014), 12–22.

[289] **Rapid7**. Phishing Awareness Training. https://www.rapid7.com/solutions/phishing-awareness-training/, 2021. Accessed: 2023-10-22.

[290] **Redmiles, E. M., Acar, Y. G., Fahl, S., and Mazurek, M. L.** A Summary of Survey Methodology Best Practices for Security and Privacy Researchers. Technical Reports of the Computer Science Department CS-TR-5055, University of Maryland, 2017.

[291] **Redmiles, E. M., Chachra, N., and Waismeyer, B.** Examining the Demand for Spam: Who Clicks? In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, Apr. 2018), CHI '18, ACM, pp. 1–10.

[292] **Redmiles, E. M., Kross, S., and Mazurek, M. L.** How Well Do My Results Generalize? Comparing Security and Privacy Survey Results from MTurk, Web, and Telephone Samples. In *Symposium on Security and Privacy* (2019), IEEE, pp. 1326–1343.

[293] **Resnik, D. B., and Finn, P. R.** Ethics and phishing experiments. *Science and engineering ethics 24*, 4 (2018), 1241–1252. Publisher: Springer.

[294] **Rezaei, S.** Beyond explicit measures in marketing research: Methods, theoretical models, and applications. *Journal of Retailing and Consumer Services 61* (July 2021), 102545.

[295] **Richardson, L.** BeautifulSoup. `https://pypi.org/project/beautifulsoup4`, 2020. Accessed: 2023-10-22.

[296] **Robertson, J. L., and Barling, J.** Greening organizations through leaders' influence on employees' pro-environmental behaviors. *Journal of organizational behavior 34*, 2 (2013), 176–194. Publisher: Wiley Online Library.

[297] **Robinson, S. L.** Trust and Breach of the Psychological Contract. *Administrative Science Quarterly 41* (1996), 574–599.

[298] **Rogers, R. W.** A Protection Motivation Theory of Fear Appeals and Attitude Change1. *The Journal of Psychology 91*, 1 (1975), 93–114. Publisher: Routledge.

[299] **Romaiha, N., Maulud, F., Musyirah, W., Jahya, A., Fahana, N., and Harun, A.** The Determinants of Organizational Citizenship Behaviour (OCB). *International Journal of Academic Research in Business and Social Sciences 9* (2019), 124–133.

[300] **Rose, S., Engel, D., Cramer, N., and Cowley, W.** Automatic Keyword Extraction from Individual Documents. In *Text Mining: Applications and Theory*. Wiley Online Library, 2010, pp. 1–20.

[301] **Rublee, E., Rabaud, V., Konolige, K., and Bradski, G.** ORB: an efficient alternative to SIFT or SURF. In *International Conference on Computer Vision* (2011), IEEE, pp. 2564–2571.

[302] **Rudoy, J., and Paller, K.** Who can you trust? Behavioral and neural differences between perceptual and memory-based influences. *Frontiers in Human Neuroscience 3* (2009).

[303] **Salahdine, F., and Kaabouch, N.** Social Engineering Attacks: A Survey. *Future Internet 11*, 4 (Apr. 2019), 89. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[304] **Salkind, N.** *Encyclopedia of research design*. SAGE, 2010.

[305] **Sanfey, A., Loewenstein, G., McClure, S., and Cohen, J.** Neuroeconomics: Cross-currents in research on decision-making. *Trends in Cognitive Sciences 10*, 3 (2006), 108–116.

[306] **Sarter, M., Givens, B., and Bruno, J. P.** The cognitive neuroscience of sustained attention: where top-down meets bottom-up. *Brain Research Reviews 35*, 2 (Apr. 2001), 146–160.

[307] **Schechter, S. E., Dhamija, R., Ozment, A., and Fischer, I.** The Emperor's New Security Indicators. In *2007 IEEE Symposium on Security and Privacy (SP '07)* (May 2007), pp. 51–65. ISSN: 2375-1207.

[308] **Schein, E.** *Organizational Culture and Leadership*. Jossey-Bass, 2010.

[309] **Scheitle, Q., Hohlfeld, O., Gamba, J., Jelten, J., Zimmermann, T., Strowes, S., and Vallina-Rodriguez, N.** A Long Way to the Top. In *Internet Measurement Conference* (2018), ACM.

[310] **Serra, J.** *Image Analysis and Mathematical Morphology*. Academic Press, 1983.

[311] **Shahbaznezhad, H., Kolini, F., and Rashidirad, M.** Employees Behavior in Phishing Attacks What Individual Organizational and Technological Factors Matter. *Journal of Computer Information Systems 61*, 6 (2020), 539–550.

[312] **Shaw, J.** Do False Memories Look Real? Evidence That People Struggle to Identify Rich False Memories of Committing Crime and Other Emotional Events. *Frontiers in Psychology 11* (2020), 650.

[313] **Shekokar, N., Shah, C., Mahajan, M., and Rachh, S.** An Ideal Approach for Detection and Prevention of Phishing Attacks. *Procedia Computer Science 49* (2015), 82–91.

[314] **Sheng, S., Lanyon, M., Kumaraguru, P., Cranor, L., and Downs, J.** Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2010), CHI '10, ACM, pp. 373–382.

[315] **Sheng, S., Wardman, B., Warner, G., Cranor, L., Hong, J., and Zhang, C.** An Empirical Analysis of Phishing Blacklists. In *Conference on Email and Anti-Spam* (2009). Publisher: Carnegie Mellon University.

[316] **Shonman, M., Li, X., Zhang, H., and Dahbura, A.** Simulating phishing email processing with instance-based learning and cognitive chunk activation. In *Brain Informatics* (2018), LNAI, Springer, pp. 468–478.

[317] **Silic, M., and Back, A.** The dark side of social networking sites: Understanding phishing risks. *Computers in Human Behavior 60* (2016), 35–43. Publisher: Elsevier.

[318] **Silic, M., and Lowry, P.** Using Design-Science Based Gamification to Improve Organizational Security Training and Compliance. *Journal of Management Information Systems 37*, 1 (2020), 129–161.

[319] **Siponen, M., Mahmood, M. A., and Pahnila, S.** Employees' adherence to information security policies: An exploratory field study. *Information & Management 51*, 2 (2014), 217–224.

[320] **Sommestad, T., and Karlzen, H.** A meta-analysis of field experiments on phishing susceptibility. In *Symposium on Electronic Crime Research* (2019), vol. 2019-November, IEEE, pp. 1–14.

[321] **Steinmetz, K. F., Pimentel, A., and Goe, W. R.** Performing social engineering: A qualitative study of information security deceptions. *Computers in Human Behavior 124* (2021), 106930.

[322] **Stembert, N., Padmos, A., Bargh, M., Choenni, S., and Jansen, F.** A Study of Preventing Email (Spear) Phishing by Enabling Human Intelligence. In *European Intelligence and Security Informatics Conference* (2016), IEEE, pp. 113–120.

[323] **Steves, M., Greene, K., and Theofanos, M.** Categorizing human phishing difficulty: a Phish Scale. *Journal of Cybersecurity 6*, 1 (Sept. 2020), 9.

[324] **Stivala, G., and Pellegrino, G.** Deceptive Previews: A Study of the Link Preview Trustworthiness in Social Platforms. In *Network and Distributed Systems Security (NDSS) Symposium* (2020), Internet Society.

[325] **Stockhardt, S., Reinheimer, B., Volkamer, M., Mayer, P., Kunz, A., Rack, P., and Lehmann, D.** Teaching Phishing-Security: Which Way is Best? In *ICT Systems Security and Privacy Protection* (2016), Springer, pp. 135–149.

[326] **Stout, B.** Email Credential Harvesting at Scale Without Malware. `https://unit42.paloaltonetworks.com/credential-harvesting/`, Sept. 2021. Accessed: 2021-12-14.

[327] **Stringhini, G., and Thonnard, O.** That Ain't You: Blocking Spearphishing Through Behavioral Modelling. In *Detection of Intrusions and Malware, and Vulnerability Assessment* (2015), pp. 78–97.

[328] **Suzuki, S., and be, K.** Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing 30*, 1 (1985), 32–46.

[329] **Sypniewska, B.** Counterproductive Work Behavior and Organizational Citizenship Behavior. *Advances in Cognitive Psychology 16* (2020), 321–328.

[330] **Tan, C. L.** Phishing Dataset for Machine Learning: Feature Evaluation. `doi.org/10.17632/h3cgnj8hft.1`, 2018. Accessed: 2023-10-22.

[331] **Tang, J., Birrell, E., and Lerner, A.** Replication: How Well Do My Results Generalize Now? The External Validity of Online Privacy and Security Surveys. In *Proceedings of the 18th Symposium on Usable Privacy and Security, SOUPS 2022* (Boston, MA, 2022), USENIX Association, pp. 367–385.

[332] **Tetri, P., and Vuorinen, J.** Dissecting social engineering. *Behaviour & Information Technology 32*, 10 (2013), 1014–1023.

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

[334] **Thomas, K., Pullman, J., Yeo, K., Raghunathan, A., Kelley, P. G., Invernizzi, L., Benko, B., Pietraszek, T., Patel, S., Boneh, D., and Bursztein, E.** Protecting accounts from credential stuffing with password breach alerting. In *USENIX Security Symposium* (2019), USENIX Association, pp. 1556–1571.

[335] **Thomas, K. A., and Clifford, S.** Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior 77* (Dec. 2017), 184–197.

[336] **ThriveDX Lucy**. 2022 Global Cybersecurity Awareness Training Study. `https://edu.nl/yrt3f`, Aug. 2022. Accessed: 2023-06-16.

[337] **Tian, Y., Yuan, J., and Yu, S.** SBPA: Social behavior based cross Social Network phishing attacks. In *2016 IEEE Conference on Communications and Network Security, CNS 2016* (2017), IEEE, pp. 366–367.

[338] **Tischer, M., Durumeric, Z., Foster, S., Duan, S., Mori, A., Bursztein, E., and Bailey, M.** Users Really Do Plug in USB Drives They Find. In *Symposium on Security & Privacy* (2016), IEEE, pp. 306–319.

[339] **Tu, H., Doupé, A., Zhao, Z., and Ahn, G.** Users Really Do Answer Telephone Scams. In *USENIX Security Symposium* (2019), USENIX Association, pp. 1327–1340.

[340] **Tversky, A., and Kahneman, D.** Judgment under Uncertainty: Heuristics and Biases. *Science 185*, 4157 (1974), 1124–1131.

[341] **Uehara, K., Mukaiyama, K., Fujita, M., Nishikawa, H., Yamamoto, T., Kawauchi, K., and Nishigaki, M.** Basic Study on Targeted E-mail Attack Method Using OSINT. In *Advanced Information Networking and Applications* (2020), Advances in Intelligent Systems and Computing, Springer, pp. 1329–1341.

[342] **Uhlmann, E. L., Leavitt, K., Menges, J. I., Koopman, J., Howe, M., and Johnson, R. E.** Getting Explicit About the Implicit: A Taxonomy of Implicit Measures and Guide for Their Use in Organizational Research. *Organizational Research Methods 15*, 4 (Oct. 2012), 553–601. Publisher: SAGE Publications Inc.

[343] **US Department of Labor**. Minimum wage. `https://www.dol.gov/agencies/whd/minimum-wage`, 2022. Accessed: 2022-5-22.

[344] **Vagias, W. M.** Likert-Type Scale Response Anchors. Tech. rep., Clemson University, 2006.

[345] **Vahdad, A.** The Human Attack Surface Framework for Phishing. Master's thesis, Eindhoven University of Technology, 2020.

[346] **Valecha, R., Gonzalez, A., Mock, J., Golob, E. J., and Raghav Rao, H.** Investigating Phishing Susceptibility—An Analysis of Neural Measures. In *Information Systems and Neuroscience* (Cham, 2020), F. D. Davis, R. Riedl, J. vom Brocke, P.-M. Léger, A. Randolph, and T. Fischer, Eds., Lecture Notes in Information Systems and Organisation, Springer, pp. 111–119.

[347] **Van Der Heijden, A., and Allodi, L.** Cognitive triaging of phishing attacks. In *USENIX Security Symposium* (2019), USENIX Association, pp. 1309–1326.

[348] **van Dooremaal, B., Burda, P., Allodi, L., and Zannone, N.** Combining Text and Visual Features to Improve the Identification of Cloned Webpages for Early Phishing Detection. In *The 16th International Conference on Availability, Reliability and Security* (New York, NY, USA, Aug. 2021), ARES 2021, ACM, pp. 1–10.

[349] **Vasek, M., and Moore, T.** Do Malware Reports Expedite Cleanup? An Experimental Study. In *USENIX Conference on Cyber Security Experimentation and Test* (2012), USENIX Association.

[350] **Verizon**. 2019 Data Breach Investigations Report. `https://enterprise.verizon.com/en-nl/resources/reports/dbir/2019/introduction/`, 2019. Accessed: 2023-04-14.

[351] **Verizon**. 2020 Data Breach Investigations Report. `https://enterprise.verizon.com/en-nl/resources/reports/dbir/2020/introduction/`, 2020. Accessed: 2023-04-14.

[352] **Verizon**. 2022 Data Breach Investigations Report. `https://www.verizon.com/business/resources/reports/dbir/2022/master-guide/`, 2022. Accessed: 2023-04-14.

[353] **Verplanken, B., Aarts, H., van Knippenberg, A., and Moonen, A.** Habit versus planned behaviour: A field experiment. *British Journal of Social Psychology 37*, 1 (1998), 111–128.

[354] **Vidas, T., Owusu, E., Wang, S., Zeng, C., Cranor, L., and Christin, N.** QRishing: The Susceptibility of Smartphone Users to QR Code Phishing Attacks. In *Financial Cryptography and Data Security* (2013), LNCS, Springer, pp. 52–69.

[355] **Vishwanath, A.** Diffusion of deception in social media: Social contagion effects and its antecedents. *Information Systems Frontiers 17*, 6 (Dec. 2015), 1353–1367.

[356] **Vishwanath, A.** Examining the Distinct Antecedents of E-Mail Habits and its Influence on the Outcomes of a Phishing Attack. *Journal of Computer-Mediated Communication 20*, 5 (2015), 570–584.

[357] **Vishwanath, A.** Habitual Facebook Use and its Impact on Getting Deceived on Social Media. *Journal of Computer-Mediated Communication 20*, 1 (Jan. 2015), 83–98. Publisher: Oxford Academic.

[358] **Vishwanath, A.** Mobile device affordance: Explicating how smartphones influence the outcome of phishing attacks. *Computers in Human Behavior 63* (2016), 198–207.

[359] **Vishwanath, A.** Getting phished on social media. *Decision Support Systems 103* (2017), 70–81.

[360] **Vishwanath, A., Harrison, B., and Ng, Y.** Suspicion, Cognition, and Automaticity Model of Phishing Susceptibility. *Communication Research 45*, 8 (2016), 1146–1166.

[361] **Vishwanath, A., Herath, T., Chen, R., Wang, J., and Rao, H. R.** Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. *Decision Support Systems 51*, 3 (2011), 576–586.

[362] **Volkamer, M., Renaud, K., and Reinheimer, B.** TORPEDO: TOoltip-poweRed Phishing Email DetectiOn. In *ICT Systems Security and Privacy Protection* (2016), Springer, pp. 161–175.

[363] **Volkamer, M., Renaud, K., Reinheimer, B., Rack, P., Ghiglieri, M., Mayer, P., Kunz, A., and Gerber, N.** Developing and Evaluating a Five Minute Phishing Awareness Video. In *Trust, Privacy and Security in Digital Business* (2018), LNCS, Springer, pp. 119–134.

[364] **Wang, J., Herath, T., Chen, R., Vishwanath, A., and Rao, H.** Research Article Phishing Susceptibility: An Investigation Into the Processing of a Targeted Spear Phishing Email. *IEEE Transactions on Professional Communication 55*, 4 (2012), 345–362.

[365] **Wang, J., Li, Y., and Rao, H.** Coping responses in phishing detection: An investigation of antecedents and consequences. *Information Systems Research 28*, 2 (2017), 378–396.

[366] **Wang, L., Zhang, Y., and Feng, J.** On the Euclidean distance of images. *IEEE Transactions on Pattern Analysis & Machine Intelligence 27*, 8 (2005), 1334–1339.

[367] **Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E.** Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing 13*, 4 (Apr. 2004), 600–612. Conference Name: IEEE Transactions on Image Processing.

[368] **Wang, Z., Sun, L., and Zhu, H.** Defining Social Engineering in Cybersecurity. *IEEE Access 8* (2020), 85094–85115. Conference Name: IEEE Access.

[369] **Wash, R., and Cooper, M.** Who provides phishing training? Facts, stories, and people like me. In *Conference on Human Factors in Computing Systems - Proceedings* (2018), vol. 2018-April, ACM, pp. 1–12.

[370] **Wenyin, L., Liu, G., Qiu, B., and Quan, X.** Antiphishing through Phishing Target Discovery. *IEEE Internet Computing 16* (2012), 52–61.

[371] **Williams, E. J., Hinds, J., and Joinson, A. N.** Exploring susceptibility to phishing in the workplace. *International Journal of Human-Computer Studies 120* (2018), 1–13. Publisher: Elsevier.

[372] **Williams, E. J., Morgan, P. L., and Joinson, A. N.** Press accept to update now: Individual differences in susceptibility to malevolent interruptions. *Decision Support Systems 96* (Apr. 2017), 119–129.

[373] **Williams, L. J., and Anderson, S. E.** Job Satisfaction and Organizational Commitment as Predictors of Organizational Citizenship and In-Role Behaviors. *Journal of Management 17*, 3 (1991), 601–617.

[374] **Winkielman, P., Berridge, K. C., and Wilbarger, J. L.** Unconscious Affective Reactions to Masked Happy Versus Angry Faces Influence Consumption Behavior and Judgments of Value. *Personality and Social Psychology Bulletin 31*, 1 (Jan. 2005), 121–135. Publisher: SAGE Publications Inc.

[375] **Workman, M.** Gaining Access with Social Engineering: An Empirical Study of the Threat. *Information Systems Security 16*, 6 (Dec. 2007), 315–331. Publisher: Taylor & Francis.

[376] **Workman, M.** A test of interventions for security threats from social engineering. *Information Management & Computer Security 16*, 5 (Jan. 2008), 463–483. Publisher: Emerald Group Publishing Limited.

[377] **Workman, M.** Wisecrackers: A theory-grounded investigation of phishing and pretext social engineering threats to information security. *Journal of the American Society for Information Science and Technology 59*, 4 (2008), 662–674.

[378] **Wright, R., Chakraborty, S., Basoglu, A., and Marett, K.** Where Did They Go Right? Understanding the Deception in Phishing Communications. *Group Decision and Negotiation 19*, 4 (2010), 391–416.

[379] **Wright, R., Jensen, M., Thatcher, J., Dinger, M., and Marett, K.** Influence techniques in phishing attacks: An examination of vulnerability and resistance. *Information Systems Research 25*, 2 (2014), 385–400.

[380] **Wright, R., and Marett, K.** The Influence of Experiential and Dispositional Factors in Phishing: An Empirical Investigation of the Deceived. *Journal of Management Information Systems 27*, 1 (2010), 273–303.

[381] **Wu, M., Miller, R., and Garfinkel, S.** Do security toolbars actually prevent phishing attacks? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2006), ACM, pp. 601–610.

[382] **Yang, L., Dumais, S. T., Bennett, P. N., and Awadallah, A. H.** Characterizing and predicting enterprise email reply behavior. In *International Conference on Research and Development in Information Retrieval* (2017), ACM, pp. 235–244.

[383] **Yang, W., Chen, J., Xiong, A., Proctor, R. W., and Li, N.** Effectiveness of a phishing warning in field settings. In *Proceedings of the 2015 Symposium and Bootcamp on the Science of Security* (New York, NY, USA, Apr. 2015), HotSoS '15, ACM, pp. 1–2.

[384] **Yang, W., Xiong, A., Chen, J., Proctor, R., and Li, N.** Use of Phishing Training to Improve Security Warning Compliance: Evidence from a Field Experiment. In *Hot Topics in Science of Security: Symposium and Bootcamp* (2017), ACM, pp. 52–61.

[385] **Ye, Q., Jiao, J., Huang, J., and Yu, H.** Text detection and restoration in natural scene images. *Journal of Visual Communication and Image Representation 18*, 6 (Dec. 2007), 504–513.

[386] **Zahedi, F., Abbasi, A., and Chen, Y.** Fake-Website Detection Tools: Identifying Elements that Promote Individuals' Use and Enhance Their Performance. *Journal of the Association for Information Systems 16*, 6 (2015).

[387] **Zhang, H., Liu, G., Chow, T., and Liu, W.** Textual and Visual Content-Based Anti-Phishing: A Bayesian Approach. *IEEE Transactions on Neural Networks 22*, 10 (2011), 1532–1546.

[388] **Zhang, H., Singh, S., Li, X., Dahbura, A., and Xie, M.** Multitasking and Monetary Incentive in a Realistic Phishing Study. In *International Human Computer Interaction Conference* (2018), BCS Learning & Development Ltd.

[389] **Zhao, M., An, B., and Kiekintveld, C.** Optimizing Personalized Email Filtering Thresholds to Mitigate Sequential Spear Phishing Attacks. In *Conference on Artificial Intelligence* (2016), AAAI Press, pp. 658–664.

[390] **Zimmermann, T.** Card-sorting: From Text to Themes. In *Perspectives on Data Science for Software Engineering*. Elsevier, 2016, pp. 137–141.

[391] **Çetin, O., Gañán, C., Altena, L., Tajalizadehkhoob, S., and van Eeten, M.** Tell Me You Fixed It: Evaluating Vulnerability Notifications via Quarantine Networks. In *European Symposium on Security and Privacy* (2019), IEEE, pp. 326–339.

Appendices

A

Appendix to Chapter 3

A.1. Parking fine phishing attack

This attack is taken from a study on phishing susceptibility [256] and it is a simple phishing attempt which pretext is a parking fine pretending to be from a local police authority. The e-mail content is reported in Listing A.1.

Listing A.1: Parking fine phishing attack from [256].

```
Our resources have indicated that you have a parking violation from
12/17/2015 at SW 89th Avenue at 3:34pm.
Please go to our website to obtain more information about the violation and
 to pay your fine or refute your ticket: <link>
```

A.2. Tailored phishing attack

This attack is taken from an experiment on tailored phishing susceptibility [61] where the authors administer treatments in randomized fashion to employees of a university and a consultancy company. The (first stage) e-mail content is reported in Listing A.2. The second stage is a replica of the organization's intranet login page hosted on a mimicked domain name.

Listing A.2: Tailored phishing attack against organizations [61]

```
From: info@{domain-name}
Subject: Your holiday hours
Dear Colleague,

To facilitate the planning of activities for the
period September to December, we invite you to provide a rough estimate of
the holiday hours you are currently planning to take until the end
of this calendar year.
Please provide this information by following this link:
{domain-name/path}

Thank you,
{signature}
```

A.3. NGO spear-phishing attack

This attack is an advanced spear phishing attack against an NGO [47]. The topic and wording is targeted to the victims, the pretext refers to real specific events that are of interest to the victims and impersonation of high-profile identities is attempted (with different techniques, like spoofing or typos, omitted in the listing). The e-mail content is reported in Listing A.3.

Listing A.3: NGO spear phishing attack from [47].

```
From: ...
Date: Mon, Mar 4, 2013 at 8:58 AM
Subject: Invitation Letter of WUC International Conference
To: ...
Dear ...,

I am writing to you from the World Uyghur Congress (WUC) and on behalf of
the Unrepresented Nations and Peoples Organization (UNPO) and the Society
for Threatened People (STP) with financial support from the National
Endowment of Democracy, cordially invites you to attend the WUC's upcoming
Conference which will be held in Geneva between 11th and 13th March 2013.

Attached you can find the invitation letter. We hope you will give a
positive consideration to this invitation, and look forward to meeting you
in Geneva. During your stay in Geneva, travel, accommodation and food are
covered by the WUC.

The WUC is a non-profit organization granted by the National Endowment for
Democracy in Washington, DC to peacufully promote human rights, democracy
and freedom for the Uyghur people in East Turkestan.
If you have any questions or queries regarding your participation, please
do not hesitate to contact me. Phone: ..., Fax: ..., e-mail: ...

sincerely,
```

A.4. LinkedIn multi-stage attack

This attack is a multi-stage, highly targeted spear-phishing attack against white-collar workers on LinkedIn [15] that actively employs collected information on its targets to forge the attack artifacts used in each stage of the attack. The LinkedIn post in Fig. A.1 refers to a (fictitious) Eliora Construction company located in the US. The offer is targeted towards a specific set of European, North African and Middle East countries where white-collar workers may be more easily appealed to it. Listing A.4 presents relevant portions of the artifacts used in the next stages of the attack.

Listing A.4: LinkedIn multi-stage attack from [15].

```
[STAGE 2]

Dear Applicant,
I write to inform you that your resume has been properly reviewed and
screened by our recruiting board and you have been found eligible for this
vacant position. Be informed that you have been shortlisted for an
interview scheduled for Friday, 12th of January 2018 at ELIORA CONSTRUCTION
 COMPANY, 1055 Metropolitan Avenue Charlotte, North Carolina, 28204, United
 States of America.
```

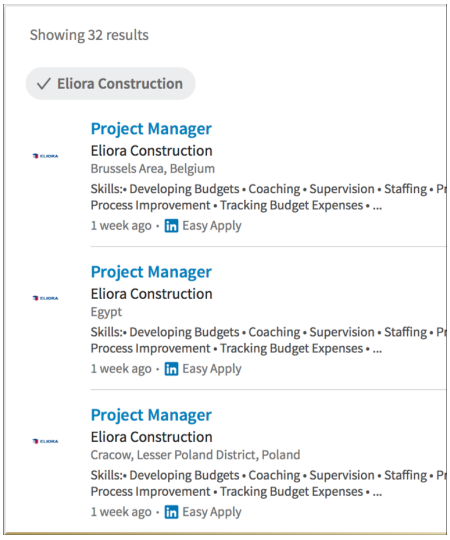


Figure A.1: Stage 1 - The LinkedIn job post

[...] our primary reason for requesting for your physical presence is to have our chief project manager have a one on one interview with you and ensure you possess the aforementioned qualities and also have you familiarize yourself with the company structure as well as a recap on past and upcoming project.

[...] Please note that our official travelling consultant shall handle your travel needs which will include flight tickets, hotel reservations, visa procurement and transfers within the United States. More so, you will be responsible for all your travel expenses made through our affiliated travel agency. These expenses shall then be refunded to you by Eliora Construction on arrival at the interview venue

[STAGE 3]

[...] Job Locations: As advertised on LinkedIn (Further information will be issued after the interview).

[...] Our company's accountant will furnish you with our banking details for making a wire transfer of your booking cost as soon as your documents have been received.

[...] Interviews are also designed to ascertain claims of working experience. Should any claim be found wanting the affected expatriate may be deported. Please note that our official travelling consultant shall handle your travel needs

[...] you will be responsible for all your travel expenses made through our affiliated travel agency.

[...] Date of Interview: Friday, 12th of January 2018

A.5. Frameworks on cognition and social engineering

Table A.1: Comparison of framework features with extant SE-related frameworks.

This framework	Other frameworks	Key aspects
Stimulus	Cranor [84]	Similarly to [57], a stimulus in [84] is a communication fed as input to the human receiver, triggering the processing steps and resulting in some behavior. The communication types in [84] can be, e.g., warnings, notices, status indicators or others, which are fully covered by stimuli and stimuli attributes in [57]. Additionally, the framework in [57] accounts for feedback across attack stages and interaction with parameters, which are not covered in [84].
Perception	Montañez et al. [238]	Perception in [238] translates signals from senses to percepts as in [57]. Perception in [57] is defined nearly identically to [238] and has additional characterizations, such as specificity of target-related information and priming, not present in [238].
Attention	Cranor [84], Montañez et al. [238]	Attention functions in [238] are described very similarly to [57] and, in both cases, concern the Working Memory (WM); [57] additionally describes the Central Executive as a voluntary attentional control system. Moreover, the frameworks [57] and [84] have an additional characterization of attention (i.e., goal-dependency in parameters [57] or attention switch [84]). Active and passive communications in [84] are related to the exogenous and endogenous attention types of [57] (i.e., as a function of certain parameters in [84]). Attention switch in [84] is captured as the routing mechanism of (exogenous or endogenous) central attention towards the heuristic or systematic processing in Elaboration in [57]. Finally, attention maintenance in [84] is the concept of sustained cognitive focus towards a warning, which is not captured in [238] and only partially captured by the routing of attention to systematic processing (e.g., a warning or indicator triggering an Anomaly) in Elaboration in [57].
Elaboration	Cranor [84], Montañez et al. [238]	Communication processing in [84] includes comprehension (to understand) and knowledge acquisition (to learn, as in training or experience), and is fully covered in [57] by the Elaboration block (reasoning towards a decision or judgment) and interactions with memory and other blocks (e.g., retrieval of experience), such as parameters and the feedback loop. Decision-making in [238] has similar functions to the Elaboration block in [57] where, in both, information is prioritized from WM to reach a decision or judgment to be translated into action. Elaboration in [57] and [84] explicitly depicts a few factors/mechanisms (e.g., Heuristics or Comprehension) relevant in the respective scopes (i.e., warning science [83] and SE attacks) that are not present in [238] (due to a higher level of abstraction of these concepts). For example, Decision-making in [238] can accommodate the usage of Heuristics or Anomalies in [57], but only Heuristics are discussed explicitly in [238]; similarly, knowledge acquisition in [84] is closely related to the interactions of [238]'s building blocks and Long-Term Memory, but with no particular focus on learning mechanisms; heuristic processing and/or cognitive biases omitted in [84], except a brief mention of habituation. Therefore, aspects related to elaboration relevant to SE attacks, such as heuristic-systematic processing, decision-making, or judgment, are fully covered only in [57].

Table A.1 continued from previous page

This framework	Other frameworks	Key aspects
Heuristics	Montañez et al. [238], MINDSPACE [98], Tetri and Vourinen [332]	The Heuristics block in [57] covers the mechanisms described in Decision-making block of [238] where decision-making, modulated by attentional mechanisms (e.g., attentional tunneling [238]) can be driven by fast and automatic processes that lead to good enough decisions, but subject to systematic errors. Heuristic's function and the stimuli attributes (e.g., usage of persuasion techniques) of [57] fully accommodate the [238]'s discussion on 'persuasion-related behavior' and are not limited to a set of persuasion techniques. The persuasion techniques and relations with certain organizational parameters in [332] are constructed similarly to the interactions between the Heuristics block, stimuli attributes, and parameters in [57]. However, the framework in [57] covers a wider set of parameters, such as personal, work, and setting-related. Similarly, the only mention of persuasion or automatic processing in [84] (i.e., the 'habituation' parameter in Communication delivery) falls within the wider characterization of heuristic processing in [57] or persuasion-related behavior of [238]. The nudging and persuasion techniques at the core of [98]'s framework fall naturally within the discussion of the Heuristics block of [57] but are almost exclusively treated standalone, in the context of policy communication, that does not relate to SE attacks. Therefore, the framework in [57] provides the widest characterization of heuristic-related processing mechanisms in the context of SE.
Anomalies	Montañez et al. [238], Cranor [84]	The Anomaly block in [57] covers the mechanisms described in Decision-making block of [238] where conscious controlled processing is defined as slower, effortful but sensitive to the particulars of a given situation (e.g. influenced by short-term factors in [238] or parameters and attributes in [57]). Knowledge transfer (Comm. processing in [84]) is captured by the Anomaly block in Elaboration in [57] and its interaction with target parameters (e.g., the ability to recognize situations and how to apply knowledge, such as with a suspicious URL in an email).
Behavior	Montañez et al. [238], Cranor [84]	Given the similarity of possible behavior types, interactions that condition the final action/not action, and assumption on the attackers, the behavior concept in [84] is fully captured by the Behavior block in [57]. Similarly, a measured action is called behavior in [238] and represents the concept of behavior as in [84] and [57]. The explicit feedback mechanism considered in [57] allows for even wider characterizations of SE attacks than the only framework that mentions it (i.e., Action in [238]).
Parameters	Cranor [84], Montañez et al. [238], MINDSPACE [98], Tetri and Vourinen [332]	The process parameters of [84] are Personal and Environmental variables (e.g., demographics, primary task), and Intentions and Capabilities (e.g., self-efficacy or conflicting goals). They define the properties and context of the information processing by the receiver. Short- and Long-term cognitive factors (e.g., workload or expertise) in [238] have a very similar role, with fewer of them being explicitly discussed. These concepts and their interactions with other processing steps (e.g., primary task w.r.t. attention type in [57]) are fully captured by parameters in [57] and their interactions with processing steps in [57] since parameters define the properties of the context of the cognitive process w.r.t. the target and environment. Some specific parameters/variables mentioned in [84] may be partially covered by [57], such as cognitive and physical skills from Capabilities in [84]. Attacker effort (i.e., message quality, personalization, and contextualization) in [238], and the relation of target and contextual variables with attacker impersonation and persuasion efforts in [332], are modeled as the alignment of target and attack parameters (attacker's assumptions on target) in [57], with no limitation on the type of parameters (virtually any property of a given communicative situation). Finally, all the references to properties of policy communication in [98] (relatable to parameters as in [84], [332] or [57]) fall naturally within the coverage of parameters in [57].

Table A.1 continued from previous page

This framework	Other frameworks	Key aspects
Features not (fully) covered		
Knowledge retention	Cranor [84]	The user's ability to remember communication when needed (knowledge retention) is made as an explicit 'processing step' in [84]. The framework in [57] loosely captures the concept by modeling the interaction of Elaboration with the Long-Term Memory and/or target parameters (e.g., knowledge within a domain of interest, such as an organization).

B

Appendix to Chapter 4

B.1. Details on analysis

After the selection of the relevant papers, the analysis was carried out by three investigators. Specifically, one investigator carried out the content analysis deductively, that is, by applying the a-priori identified criteria in Section 4.3.3 (henceforth, we call them features) on the body of the included articles. Each article has a relatively constant structure that includes at least a section with hypotheses or RQs, a study design description (usually in the methods section), and a results section. This makes it straightforward to identify the relevant information that belongs to a feature as, in most cases, the presence of a unit of information (e.g., population or stimuli) and its employment (e.g., treating participants with training or measuring clicks on links) is evident from the article text as is, with little to no risk of different interpretations [228]. To assist with the analysis, the Zotero reader and bibliography manager was used to catalogue papers and highlight relevant sections with notes for future reference; an online spreadsheet was used to build the dataset, the codebook with descriptions of features and categories and to keep track of updates, with comments and additional notes.

To support the analysis and ensure its reproducibility, one investigator defined a codebook based on the cognitive framework presented in Chapter 3, including a description of the inclusion criteria for each feature (with examples) and individual variables extracted from the papers. For example, stimuli attributes are "*Features of stimuli defining its content and form used as experiment variables.*". An example of stimuli attribute is 'framing' [145], which belongs to the category of persuasion techniques and is defined as "*The employed persuasion techniques/principles/wording or indicators to intentionally trigger cognitive biases/heuristics*". Variables concerning attributes, target parameters, and elaboration have been grouped into categories due to the sheer number of different variables reported in the papers. These variables have been grouped into categories by affinity using a bottom-up approach. It is worth noting that some categories across personal, work-related, and setting-related target parameters have the same name (i.e., demographics, experience, situation, and security awareness) although they refer to different target parameters, within which they should be interpreted. For example, 'demographics' in work-related parameters contains demographic characteristics related to the work domain of a person such as salary, while 'demographics' in personal

parameters contains variables such as age or education.

To ensure the quality of the analysis, the three investigators regularly carried out meetings (either weekly or biweekly, depending on progress) throughout the study in which they discussed the coded papers, identified points of disagreements and resolved disagreements, updating the codebook and the analysis of papers already coded [228]. To assure clarity of the descriptions and to validate the feature inclusion criteria and the categories, co-coding sessions with all three investigators were carried out whereby a random sample of papers (from three to five per session) was assigned to each investigator to independently perform the mapping of paper contents to the features and to the relative categories following the codebook. This procedure was carried out iteratively, in in-person sessions of at least 2 hours with distinct samples of papers in each session. Initially, most disagreements were caused by the ambiguous description of certain features in the articles which allowed different interpretations by readers following the initial codebook. However, with the progress of the analysis, disagreements decreased due to the updated codebook definitions lowering the chances of ambiguous interpretations.

As mentioned in threats to validity (Section 4.5.1), the mapping of extracted variables to some of the features necessarily remain subject to some level of interpretation. Somewhat ambiguous cases concerned, for example, the application of individual targetization level; specifically, whether messages saluting the recipient by name or username should be labeled as individual-level targetization. It was decided to not label such cases as individual-level, since the name or username can be trivially derived from, e.g., the email address itself. Another case of disambiguation concerned the mapping of effects on heuristics and anomalies: although studies may refer to investigations of, e.g., activation of heuristics or detection of anomalies, most of the time the measurements of these effects are carried out on the final behavior as a proxy variable, with a prior manipulation of stimulus attributes (e.g., a pretext reflecting authority or a warning pop-up). Therefore, it was decided to label such occurrences of measurements as *indirect*. Similarly, we label indirect effect measurements for attention and perception.

The codebook, description of categories, and the full dataset are available as supplementary material at `https://zenodo.org/record/8380243`.

B.2. Analysis of sample size

Fig. B.1 shows the distributions of sample sizes for each study type with extremes and median highlighted in red. The typical laboratory experiment includes less than one hundred participants[1], whereas interviews comprise around thirty participants. On the contrary, field experiments and surveys are almost always carried out in the hundreds or more, and observational studies in the tens of thousands, although this is limited to three studies only.

[1]Some lab experiments are conducted online, e.g., on Amazon Mechanical Turk, which allows recruiting a higher number of subjects compared to in-person studies.

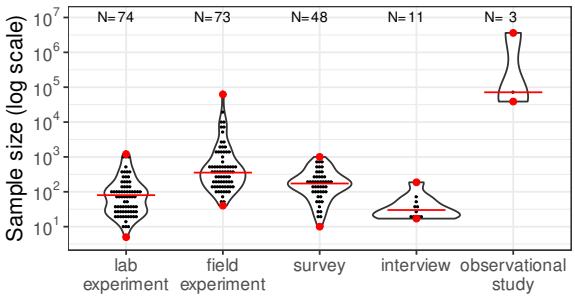


Figure B.1: Distribution of sample sizes across study types.

C

Appendix to Chapter 8

C.1. Phishing email pretext

The pre-text regarding [Trips Website][1] makes a distinction between employees who have worked at **CompanyX** for less than a year and those who have worked for more than a year.

C.1.1. Baseline (non-personalized) and Treated (personalized) emails

Dear <*EmployeeName*>,
Since the COVID measures in the Netherlands have been largely abolished, we can finally get back to our normal lives.
To let you know how much we appreciate having you at **CompanyX**, we would like to offer you the following; We partnered with [Trips Website].NL to arrange a selection of exciting activities. You can refer to the following website to view activities' options and make a reservation.
<LinkToFakeWebsite/wattedoen>
We hope that you like the idea and enjoy the activities of your choice.
– –
Kind regards/Met vriendelijke groet ,
[CEO's name]
CEO
[Company's name]
[Company's address]
[CEO's phone number]
[Company's website]

The treatment used for the employee having worked at **CompanyX** for more than a year in blue and the one used for the employee at **CompanyX** with less than than a year in red.

[1] An online platform that is well-known among Dutch audience, and specializes in offering various activities and tourist destinations in the Netherlands

Dear <*EmployeeName*>,
Since the COVID measures in the Netherlands have been largely abolished, we can finally get back to our normal lives.
To let you know how much we appreciate your long-standing efforts with us / Given that you have recently joined us and to welcome you on board once again, we would like to offer you the following; We partnered with [Trips Website].NL to arrange a selection of exciting activities in your city, <*CityName*>. You can refer to the following website to view activities' options and make a reservation.
<LinkToFakeWebsite/wattedoen>
We hope that you like the idea and enjoy the activities of your choice.
– –
Kind regards/Met vriendelijke groet ,
[CEO's name]
CEO
[Company's name]
[Company's address]
[CEO's phone number]
[Company's website]

C.2. Debriefing

The debriefing page informs victims about the experiment (i.e., purpose, authors, which data have been collected and how it will be used, and experiment authorization) and provides the authors' contact information for any questions or concerns about the experiment. To ensure experiment integrity, the debriefing page also invites participants to refrain from disclosing information about the experiment to their colleagues. Upon completion of the experiment, we provided debriefing to those participants who did not fall for the phishing attempt, regarding the nature and purpose of the campaign. The content of the debriefing page follows:

Phishing exercise DISCLAIMER
Please read the full text

Hi there, this was a phishing exercise conducted by **CompanyX** in collaboration with security researchers at **the Research Institute**. The data used in the emails sent to you was collected using public sources, such as LinkedIn and Facebook. If you have any questions, please contact **names of the persons responsible for the experiment at CompanyX and at the Research Institute.**

The password you entered has not been sent to the server and the fact that you entered your login details will not be communicated to anyone, not even your company. All data is anonymized, but we do recommend that you consider changing your password if you have not done so in a while.

It is extremely important that you do not inform your colleagues about this exercise, as it will affect the results of the experiment. The experiment will conclude on the 17th of June, from then on feel free to discuss it with your colleagues. A survey will follow, which you are kindly requested to fill out as part of the research project, to better understand the interaction between the subject and the phishing email.

More details about this experiment will be communicated to you soon.

C.3. Interview questions

1. The security awareness of the employees
    - How would you rate your overall knowledge regarding security awareness and specifically phishing? Not knowledgeable at all / somewhat knowledgeable / very knowledgeable / expert
    - Where does your knowledge stem from? Training / professional experience / previous victimization / something else?
2. The rational and emotional response upon reading the email.

- *After sharing the screen to display the email sent and highlighting certain aspects of it, such as that the sender pretends to be their CEO, pretext, and personalization.* Did you see the email before or after being warned by your colleagues?
- What did you think when you first read the email? What did you do about it?
- How did you feel when you read the email? Why do you think you felt that way?

3. The emotional drive that led employees to report the phishing campaign.
   - Did you report the phishing email? Why did you report it? / why did you not report it?
   - Are you aware of any security reporting mechanism at **CompanyX**? How would you have behaved if there was one?
4. The behavior fostered by the tailored nature of the attack.
   (a) Would you have behaved identically to other generic phishing emails or were there any features in this email that triggered your specific behavior?
   (b) Considering the usage of personal data in the email (*Place of Residence* and *Years in Current Company*), did you assume you were the only target? How did that make you feel?

D

Appendix to Chapter 9

Table D.1: Construct correlations between human factors and cyber security behaviors from previous studies

Construct	1	2	3	4	5	6	7	8	9	10
1. Civic Virtue	-									
2. Leader Support	-.08 [203]	-								
3. Organizational Commitment	.03 [203]	-	-							
4. Sportsmanship	.045** [203]	-.02 [203]	-.04 [203]	-						
5. Conscientiousness (OCB)	.035** [203]	-.03 [203]	.01 [203]	.019 [203]	-					
6. Job satisfaction	.01 [203]	-	.44** [284]	.03 [203]	.00 [203]	-				
7. Altruism	.023** [203]	.35** [203]	.21** [203]	.048** [203]	.037** [203]	.23** [203]	-			
8. Courtesy	.012** [203]	-.02 [203]	-.06 [203]	.046** [203]	.023** [203]	.02 [203]	.031** [203]	-		
9. Agreeableness	-	-	-	-	-	-	-	-	-	
10. Conscientiousness (Big Five)	-	-	-	-	-	-	-	-	.39*** [100]	-
11. Openness to Experience	-	-	-	-	-	-	-	-	.35*** [100]	.33*** [100]
12. Extraversion	-	-	-	-	-	-	-	-	.30*** [100]	.11* [100]
13. Emotional Stability	-	-	-	-	-	-	-	-	-	-
14. (Email) Habits	-	-	-	-	-	-	-	-	-	-
15. Subjective Norms	-	-	-	-	-	-	-	-	-	-
16. Self-efficacy	-	-	-	-	-	-	-	-	-	-
17. OCBO	-	.03 [203]	-.02 [203]	-	.03 [203]	.01 [203]	.42*** [100]	-	.42*** [100]	.48*** [100]
18. OCBI	-	.30** [203]	.21** [203]	-	.22** [203]	.24** [203]	.36*** [100]	-	.36*** [100]	.35*** [100]
19. OCB	.37 [251]	.41** [203]	.32** [203]	.51** [251]	.13 [203]	.31** [203]	-	-	-	-
20. SAB	-	-	-	-	-	-	-	-	.15** [100]	.18** [100]
21. SCB	-	-	-	-	-	-	-	-	.29*** [100]	.27*** [100]

with: $^{*}p \leq .05$, $^{**}p \leq .01$, $^{***}p \leq .001$

Table D.2: Continuation of Table D.1

Construct	11	12	13	14	15	16	17	18	19	20	21
1. Civic Virtue											
2. Leader Support											
3. Organizational Commitment											
4. Sportsmanship											
5. Conscientiousness (OCB)											
6. Job satisfaction											
7. Altruism											
8. Courtesy											
9. Agreeableness											
10. Conscientiousness (Big Five)											
11. Openness to Experience	-										
12. Extraversion	.22*** [100]	-									
13. Emotional Stability	-	-	-								
14. (Email) Habits	-	-	-	-							
15. Subjective Norms	-	-	-	.76 [311]	-						
16. Self-efficacy	-	-	-	.77 [311]	.59 [311]	-					
17. OCBO	.29*** [100]	-.02 [100]	.27*** [100]	-	-	-	-				
18. OCBI	.29*** [100]	.13** [100]	.08 [100]	-	-	-	.34*** [100]	-			
19. OCB	-	-	-	-	-	-	-	-	-		
20. SAB	.11* [100]	.07 [100]	.19*** [100]	-	-	-	.15*** [100]	.18*** [100]	-	-	
21. SCB	.27*** [100]	.03 [100]	.21*** [100]	-	-	-	.44*** [100]	.22*** [100]	-	.25*** [100]	-

D.1. Construct correlation values extracted from literature

Tables D.1 and D.2 report the construct correlation values from the literature.

D.2. Sample size calculations

We performed the Fisher's Exact Test [118] on the pilot data to observe which controls, from the selected list of eight controls, needed to be considered for calculating the sample size. The resulted p-values from this test, assessing the statistical significance of each control with the intention to report, are presented in Table D.3. We considered controls with $p \leq 0.05$ as statistically *significant*, and p-values higher than 0.05 but $p \leq 0.1$ as *borderline significant*. Therefore, the selected controls, with a maximum p-value of 0.052, were: Education ($p = 0.002$), Phishing victim ($p = 0.003$), Current employment position ($p = 0.052$), and Reporting frequency ($p =< 0.001$). Table D.4 presents the calculated parameters, as well as the computed minimum sample size, predicted from the data gathered from the pilot. The resulting maximum value for the minimum sample size was $n = 267$.

Table D.3: p-value of Controls in relationship with the Intention to Report

Control	Intention to report: p-value
Gender	0.397
Age	0.852
Education	0.002
Current occupation	0.330
Current employment position	0.052
Current employment duration	0.388
Phishing victim	0.003
Reporting frequency	< 0.001

Table D.4: Sample size calculation

Exposure	**Outcome Prevalence**		
	Intention to report phishing pO = 62 %		
		Exp. PR 1.50	Exp. PR 2.00
Education	**Power**	**PONE: 45%**	**PONE: 35.3%**
College/Univ.:75.8% (E)	80%	n = 200	n = 82
Other: 24.2% (NE)	90%	n = 267	n = 109
r: 0.32			
Employment position	**Power**	**PONE: 50.5%**	**PONE: 42.6%**
Manag./Sen. manag.:45.5% (E)	80%	n = 118	n = 42
Other: 54.5% (NE)	90%	n = 158	n = 57
r: 1.2			
Phishing victim	**Power**	**PONE: 55.3%**	**PONE: 49.9%**
Yes: 24.2% (E)	80%	n = 132	n = 41
No: 75.8% (NE)	90%	n = 177	n = 56
r: 3.13			
Reporting frequency	**Power**	**PONE: 57.9%**	**PONE: 54.3%**
Always: 14.1% (E)	80%	n = 183	n = ND
Other: 85.9% (NE)	90%	n = 245	n = ND
r: 6.1			

Note: ND = value could not be determined, as prevalence of outcome in the exposed would be above 100%, according to the specified parameters.

D.3. Questionnaire

Table D.5 presents the questions to gather participants' demographic information, and Table D.6 presents the survey item. For each constructs, the latter table reports the hypotheses and the corresponding characteristics, along with which survey items measure the selected characteristics, and the study which served as reference.

Table D.5: Demographic questions in the survey

No.	Demographic Question	Answer Options
C1	What is your gender?	Male Female Prefer not to say Other
C2	What is your age in years?	Young adult (18–30) Adult (31–50) Senior adult (> 50) Prefer not to say
C3	What is the highest degree or level of school you have completed? If currently enrolled, please select the highest degree you have already completed.	Primary School Secondary/High School College/University
C4	Which of the following categories best describes your current position, if any? Note: If 'Not employed' or 'Retired' is selected, please consider your affiliation with the last organization when answering the upcoming questions.	Student Employed/Self-employed Not employed Retired Other, please specify
C5	Which of the following categories best describes your employment position/role at the organization you are affiliated with?	Intern Entry-level/Associate Manager/Senior manager C-level executive/Director/Owner Other, please specify
C6	For how long have you been in your selected position regarding the previously mentioned affiliation with the organization?	Less than half a year Between half a year and 2 years More than 2 years
C7	As far as you know, have you ever fallen for a fraudulent phishing email?	Yes No
C8	When you receive an email in your inbox that you consider suspicious, how often do you report it?	Never Rarely Occasionally Frequently Always

Table D.6: Survey items

Hypothesis	Characteristic	Survey Item	Reference
Part I			
H1.5, H3.5	Emotional Stability	At my workplace… *(ES1)* I am relaxed most of the time. *(ES2)* I often feel sad/discouraged. (R) *(ES3)* I get stressed out easily. (R) *(ES4)* I worry about things. (R)	[126]
H1.6, H3.6	Extraversion	At my workplace… *(E1)* I feel comfortable around my co-workers. *(E2)* I do not mind being the center of attention. *(E3)* I do not talk a lot with my co-workers. (R)	[126]

…continued

Hypothesis	Characteristic	Survey Item	Reference
		(E4) I do not like to draw attention to myself. (R)	
H1.1, H3.1	Sportsmanship	*(S1)* I spend a lot of time complaining about trivial matters to my co-workers. (R) *(S2)* I always focus on what is wrong at work, rather than the positive side. (R) *(S3)* I tend to make problems seem worse than they actually are. (R) *(S4)* I criticize/find fault in what the organization is doing. (R)	[283]
H1.2, H3.2	Conscientiousness	*(CNS1)* I treat my punctuality at work with seriousness. *(CNS2)* I take no undeserved breaks at work. *(CNS3)* I follow the organization's informal rules and policies, even when no one is watching. *(CNS4)* I am committed to diligently putting in the amount of work expected by my employer.	[283]
H1.3, H3.3	Altruism	*(A1)* I assist my co-workers with their tasks when they have been absent or have heavy workloads. *(A2)* I go out of my way to help new co-workers within the organization. *(A3)* I willingly lend a compassionate ear to co-workers who have work-related or personal problems. *(A4)* I willingly lend a helping hand to the co-workers around me when they need me.	[283]
H1.4, H3.4	Courtesy	*(CO1)* I take steps to try and prevent creating problems for other employees (i.e., changing holiday schedule / work days / shifts). *(CO2)* I am mindful of how my behavior affects my co-workers' jobs. *(CO3)* I do not abuse the rights of my co-workers. *(CO4)* I consider the impact of my actions on my co-workers.	[283]
Part II			
H3.7	Self-efficacy	*Scenario:* You are part of Western University, an institution which encourages employees to follow their strongly defined data policies and regulations that aim at protecting the organization's private data. John works as an HR advisor within the Human Resources Management department of the Western University. This university has a strong Information Security Policy that requires stringent compliance with email security requirements. This policy requires that suspicious emails must be reported to the Information Security department of the university. Due to this role in the university, John sends and receives numerous emails on a regular basis from job agencies, as well as from possible job candidates. One such received email from a trusted job agency contained cues that made John suspicious that the email could be a phishing email. He contacted the job agency in order to warn them about the possibility of them being impersonated by an attacker in a phishing incident. However, considering that he could recognize the email as phishing, John did not value it as a high-risk threat. Therefore, he did not report it to the university's Information Security department, and simply deleted it from his inbox. *(SE1)* I am confident that if I find myself in John's position, I would be able to contact the job agency about the suspicious email. *(SE2)* I am confident that I am able to report by myself an email that I found to be suspicious. *(SE3)* At this moment, I am confident that I am able to report an email I find suspicious, even if there was no one around to tell me what to do. *(SE4)* At this moment, I am confident that I am able to report an email I find suspicious, if I could ask for help when I am stuck.	[311]
H3.8	Subjective Norms	I believe that… *(SN1)* my supervisors think that I should put effort into protecting the private data of the organization. *(SN2)* my colleagues think that protecting the private data of the organization is our responsibility.	[311]

…continued

Hypothesis	Characteristic	Survey Item	Reference
		(SN3) my supervisors think that I should increase my performance at work, and to do so, I overlook/omit the obligations I have for protecting the private data of the organization. (R)	
		(SN4) my organization's IT department thinks that I must follow the Information Security policies.	
H2.1	Positive Cyber	*(POS1)* I monitor my work computer for signs of a virus and/or malware.	[100]
	Security Behaviors	*(POS2)* I immediately report suspicious emails I receive at work after reading them.	
		(POS3) I go above and beyond what is required of me in order to protect the private data of the organization.	
		(POS4) I follow the Information Security Policies and practices of the organization I work for.	
		(POS5) I use the Information Security technology provided to me by the organization I work for.	
		(POS6) I comply with organizational Information Security Policies in order to protect the organization's Information Systems.	
Part III			
	Intention to Report	I believe that reporting suspicious emails to the organization's IT department may be desirable because…	
	Phishing Emails	*(REP1)* it is required by the email security policy of the organization I work for.	
		(REP2) it is important to contribute to protecting the organization that I work for as a whole.	
		(REP3) it is important to contribute to protecting the information and technology resources of the organization I work for.	
		(REP4) it is important to contribute to protecting my colleagues from similar attacks.	
		Open question: Is there any remark you would like to make on why anybody, you included, may or may not want to report phishing emails?	-
	Attention check:	This is an attention check question, so please click on the answer 'Occasionally'.	

Notes: R = reverse scored question

D.4. Regression analysis results

Table D.7: Linear Regression Results

Group	Factor	Model 1 DV: $PCSB(Ui)$		Model 2 DV: $RepInt(Ui)$		Model 3 DV: $RepInt(Ui)$	
OCBO	**Sportsmanship**	-.082	.021			-.143*	-.169*
	Conscientiousness	.386***	.359***			.070	.089
OCBI	**Altruism**	.160*	.096			.187**	.174**
	Courtesy	.145†	.088			.091	.071
Pers. Attr.	**Emotional Stability**	-.067	-.075			.120†	.125†
	Extraversion	.183**	.122*			-.050	-.054
Beliefs	**Self-efficacy**					.348***	.325***
	Subjective Norms					.240***	.207***
	Positive Cyber Sec Behaviors			.651***	.630***		
Controls	**C1 (Gender)**		.030		-.055		-.031
	C2 (Age)		-.025		.030		.042
	C3 (Education)		.027		.001		-.008
	C4 (Current occupation)		-.055		.044		.013
	C5 (Current empl. position)		.037		-.119**		-.090*
	C6 (Current empl. duration)		.113*		.036		.047
	C7 (Phishing victim)		.056		-.127*		-.027
	C8 (Reporting frequency)		.377***		.058		.131**
	Adjusted R^2	.378	.521	.422	.455	.535	.549
	F	29.636	22.959	207.765	27.207	41.762	22.504
	$Obs.$	284	284	284	284	284	284

†$p \leq .1$
*$p \leq .05$
**$p \leq .01$
***$p \leq .001$

Curriculum Vitæ

Pavlo was born in Horodok, Ukraine. He studied computer engineering at the University of Pavia in Italy where he obtained both a BSc (in 2015) and an MSc degree (in 2017).

Apart from that, he studied for one semester at Eindhoven University of Technology, Netherlands as an exchange student. During his studies, Pavlo worked as student assistant for several courses at the University of Pavia as well as a lecturer at the Alessandro Volta college in Pavia. During his Master thesis, he worked as a security analyst for a managed security services company in Milan, Italy.

From 2018 he started a PhD-TA (PhD-Teaching Assistant) project at the Eindhoven University of Technology where he gained knowledge about human factors for information security of which the results are presented in this dissertation. The PhD-TA track includes additional teaching tasks, such as instructions and lectures. As a member of the MC&S PhD Council, Pavlo assisted council activities, such as organizing social events and coordinating communication within the TA program to improve the experience of his fellow PhD-TAs.