Cognition in Social Engineering Empirical Research: a Systematic Literature Review

PAVLO BURDA, Eindhoven University of Technology, Netherlands
LUCA ALLODI, Eindhoven University of Technology, Netherlands
NICOLA ZANNONE, Eindhoven University of Technology, Netherlands

The interdisciplinarity of the Social Engineering (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. In this work, 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. 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 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. Our findings on current SE research dynamics provide insights on methodological shortcomings and help identify supplementary techniques that can open promising future research directions.

CCS Concepts: • **Security and privacy → Social engineering attacks**; • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: Social Engineering, cognitive processes, empirical studies

ACM Reference Format:
Pavlo Burda, Luca Allodi, and Nicola Zannone. 2022. Cognition in Social Engineering Empirical Research: a Systematic Literature Review. *ACM Comput. Surv.* 1, 1 (April 2022), 35 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Humans are a critical component of any computer system and, as such, are part of a system's attack surface. Social Engineering (SE) attacks aim to deceive individuals with the goal of gaining access to sensitive information (such as the target's credentials) and/or deliver malware on the target's system (e.g., to "recruit" it into a botnet) [119]. The 'human vulnerabilities' exploited by SE attacks are ingrained in human cognition and, thus, can be exploited by an attacker by providing 'input' to a legitimate user of the system. As these vulnerabilities are shared across 'human targets', they represent a rather stable attack surface, and allow attackers to avoid the complexity and costs associated with deploying malware-based attacks [5, 79, 122]. For these reasons, SE attacks have gained a prominent role in the current threat

Authors' addresses: Pavlo Burda, p.burda@tue.nl, Eindhoven University of Technology, Netherlands; Luca Allodi, Eindhoven University of Technology, Netherlands, l.allodi@tue.nl; Nicola Zannone, Eindhoven University of Technology, Netherlands, n.zannone@tue.nl.

landscape of information security [177, 178] and became the most common attack vector targeting Internet users at large [26, 178], as well as enabling high-profile subversive attacks aiming at disrupting democratic processes [12, 21, 58].

Recent studies have 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 (of the '*your email password will expire in 24hrs, click here to not lose all your emails*' type), into tailored multi-step attacks leveraging target weaknesses and target-specific information [6, 19]. Recent examples include lateral movement attacks in organizations to study correspondence habits [80], multi-stage attacks bypassing two-factor authentication at scale [75] and the continuous adaptation of attacks to societal conditions, such as the COVID-19 pandemic [25] or the widespread usage of QR codes as an attack vector [42]. These attacks are often tailored against specific organizations or groups of people, exploiting the specificity and characteristics of their targets [26, 80]. 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. Stemming from this, a new strain of empirical research emerged investigating the interplay between attack features and cognitive effects [122, 144, 162]. However, the interdisciplinarity of the SE domain makes it particularly difficult to identify gaps and open research questions as well as to interpret experimental results [6, 32, 79, 167, 189]. 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 work, we carry out 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 an established cognitive framework of SE [32] 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 [6] – *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 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 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.

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 [144] survey behavioral science findings relevant to cyber-security, which partially cover cognitive process features, for example, elaboration and behavior. Darwish et al. [44] 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

Table 1. Literature studies on Social engineering with their coverage (● means "covered", ◐ "partially covered", ○ "not covered").

		Pfleeger & Caputo [144]	Darwish et al. [44]	Heartfield & Louka [79]	Purkait [148]	Tetri & Vourinen [167]	Montañez et al. [122]	Sommestand & Karlzen [162]	Salahdine & Kaabouch [153]	Franz et al. [64]	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	○	○	○	○	○	○	◐	○	◐	◐

Loukas [79] propose a taxonomy of semantic SE attacks along with their characteristics and review defense techniques, similarly to Salahdine and Kaabouch [153] and Purkait [148]. Tetri and Vourinen [167] introduce a conceptual framework for SE to analyze attack characteristics, parameters of targets and setting, and the execution SE attacks. Montañez et al. [122] 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 [162] 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. [64] present a taxonomy of phishing interventions for usable security that comprehends the design of training experiments, including 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 1 presents a comparison between our literature review and other surveys. Only the survey in [162] 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 3.3 for the motivations underlying this choice). Other literature reviews [122, 144, 167] 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 [64] relates parts of cognitive processes and experiment design characteristics to SE attacks, but focus mainly on prevention and user interaction aspects. The other reported surveys (i.e., [44, 79, 148, 153]) do not treat cognitive-related aspects neither look into the experiment design.

Outline. The paper is structured as follows. The next section introduces the background concepts on social engineering, cognitive processes and empirical approaches adopted by SE studies relevant for the analysis. Section 3 describes our methodology for data collection, lays out the research questions and derive the criteria used in the analysis. Section 4 presents the results of the analysis and Section 5 discusses our findings. Finally, Section 6 concludes the paper and provides actionable insights and future directions.

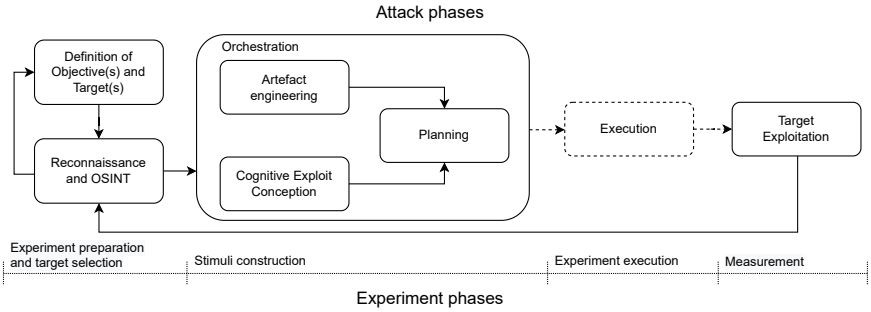

Fig. 1. The schema of SE attack (top) and experiment (bottom) phases.

2 BACKGROUND

To develop a structured overview of the empirical SE research and cognition we introduce background notions on SE attack phases and characteristics, the cognitive processes at play during target exploitation, as well as the empirical approaches adopted in the SE literature.

2.1 Social Engineering

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 [189]. Fig. 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 [6]. 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 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 [6, 26], in stark contrast with 'classical' SE attacks that are untargeted in nature and employ simple techniques to persuade their victims [145]. 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 collects 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 their final objective [6, 80].

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), to lure targets and elicit information [172, 182]. 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 [30, 171, 179]. Examples are tailgating or dissemination of malicious QR codes, USB drives and decoy wireless access points. SE attacks play also 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 [26, 107].

Regardless of the level of sophistication, SE attacks tend to exploit 'vulnerabilities' inborn in human cognition, e.g. faulty beliefs and cognitive patterns [34, 61, 95, 153, 167]. 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 (that being 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 [198]) and the involved processes (such as priming subjects before deploying an attack [95]), must be considered to understand the underpinning mechanisms that bring to victimization. In the next section, we present the main components over which the relevant cognitive processes develop.

2.2 Cognitive Processes

Cognitive sciences have identified a general set of components that constitute the architecture of human cognitive processes, particularly in the fields of psychology, linguistics and neuroscience [14, 78]. This results in a variety of theories and models accounting for mental capabilities of perception, memory, attention, reasoning, etc. However, the connection between theories of cognition and SE attacks is less widely explored. For the purpose of this work, we adopt the cognitive framework for SE proposed in [32]. This framework is distilled from existing theories of cognition and aims to provide a means to structure and analyze SE attacks from a cognitive perspective. An overview of the framework and its building blocks is shown in Fig. 2; the building blocks and their relevance for SE attacks are further detailed in Table 2.

A cognitive process is triggered by the arrival of a *stimulus* (in the context of SE, for example, an email or a phone call). Cognition translates the stimulus into *percepts* (*perception*, where concepts and procedures about email communication, awareness, writing style etc. are pre-loaded in working memory) [14, 16, 109] and routes those via *attention* (e.g., given the relevancy of that email) to the *elaboration* system where the information contained in the stimulus is processed (e.g., weighting different factors from the stimulus with contextual information and percepts to formulate a decision whether to ignore or respond to that email) [16, 18, 45, 59]. A *behavior* is produced as the output of the cognitive process (e.g., clicking a link, deleting or reporting the message, opening an attachment). This process can be mediated by *Parameters* characterizing the subject and their context (in the instant the stimulus is processed), and influencing the cognitive process at all levels (perception, attention, and elaboration) [47]. Parameters can be *attack parameters* (α), denoting the (explicit or implicit) assumptions of the attacker about the targeted subject characteristics and context, and *target parameters* (θ), which denote the *actual* characteristics and context of the target. Examples of *parameters* can be short-term influencing factors, such as the current work load, or long-term, such as personality traits [122]. Attack and target parameters are further grouped into three main dimensions: personal subject dimension (α^p, θ^p), subject's work-related/main activity dimension (α^w, θ^w) and subject's setting dimension (α^s, θ^s); for details, refer to [32].

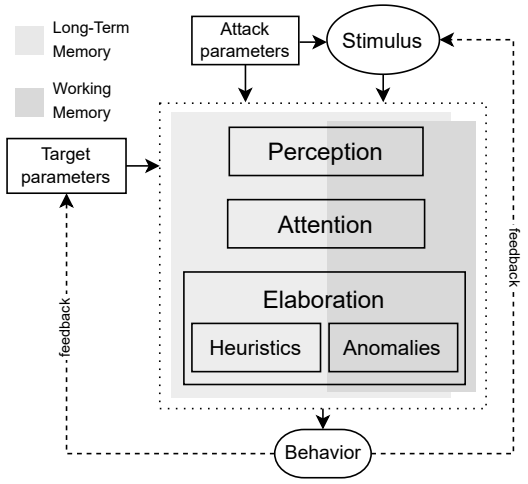

Fig. 2. Conceptual framework of cognition for SE attacks (cf. Table 2 for the blocks' description).

The described cognitive process runs on top of the Long-Term Memory, Working Memory and Central Executive modules substrates, that serve as resources and controllers for the building blocks [14, 17, 18, 78]. The interested reader can refer to [32] for further details on the framework, and examples of its application to real-world SE attacks.

Table 2. Overview of the building blocks for cognition and social engineering. Full details are reported in [32]

Component	Description	Relevance for SE
Stimuli	Any input (e.g., an event, a sound, a message) that triggers a cognitive process. A stimulus is characterized by attributes describing its content and form.	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 [116], style of writing [114], or the presence of text aimed to evoke past memories of the target [199].
Perception	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 2) but may have strong effects on the subject's subsequent decisions [40, 95]
Attention	A set of systems that modulate the access to consciousness. Here attention refers to the central attention which 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, which has been demonstrated to lead to higher deception rates [125].
Elaboration	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. `Working Memory` 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 [61, 76, 186].
Anomaly	A condition when *Elaboration* is unable to handle information that does not fit an automated processing pattern due to, e.g., wrong or lack of contextual cues, and engages in effortful processing, like consciously directing attention and making use of `Working Memory`.	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) [103].
Heuristic	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 [194] or social networks [184], and are thought to significantly affect the success of attacks [33, 134, 198].
Behavior	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. It can produce a new stimulus and initiate a new cognitive cycle in a feedback fashion.	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. We distinguish between *attack parameters* (α) and *target parameters* (θ): α represents the assumptions that the attacker makes on the targets and their context; θ characterizes the properties of the target and the context in which the target is when the external stimulus arrives.	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 [65, 69].

Substrate	Description
`Long-Term Memory`	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`	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`	An attentional control system that voluntarily manipulates the `Working Memory` functions.

2.3 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 [103], 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 [163], or factors related to cognitive processes, such as participants' gaze [120]. 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 [7], where any independent variable is out of control of the investigators. Other types of experiments commonly adopted in empirical SE research are *surveys* and *interviews*, either by themselves, e.g. [51], or complimentary to lab and field experiments [197]. These experiment types are particularly relevant when an effect cannot be observed directly, for example the rationale behind the detection of an anomaly [197] or decision making [185].

The process to setup an experiment largely matches that of real attacks (cf. Fig. 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 relevant 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 encompasses the reconnaissance phase of a real attack, in which the attacker identifies targets of interest, as shown in Fig. 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. 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 artefacts may be adapted to the subjects and include cognitive exploits or other features. The *adaptation* of the artefacts 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 [6, 34] and relative to those of non-tailored attacks [35, 63]. 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 link, credential submissions) or other indirect effects.

3 SYSTEMATIC LITERATURE REVIEW PROCESS

This section presents the methodology used to conduct the literature review.

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:

RQ *What are the open gaps and promising future directions in the empirical SE field?*

We refine this question 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. 3.

Empirical approaches adopted by SE studies: We first explore the various empirical methods adopted in the literature to understand how researches reproduce the SE attack process described in Section 2 and the different aspects

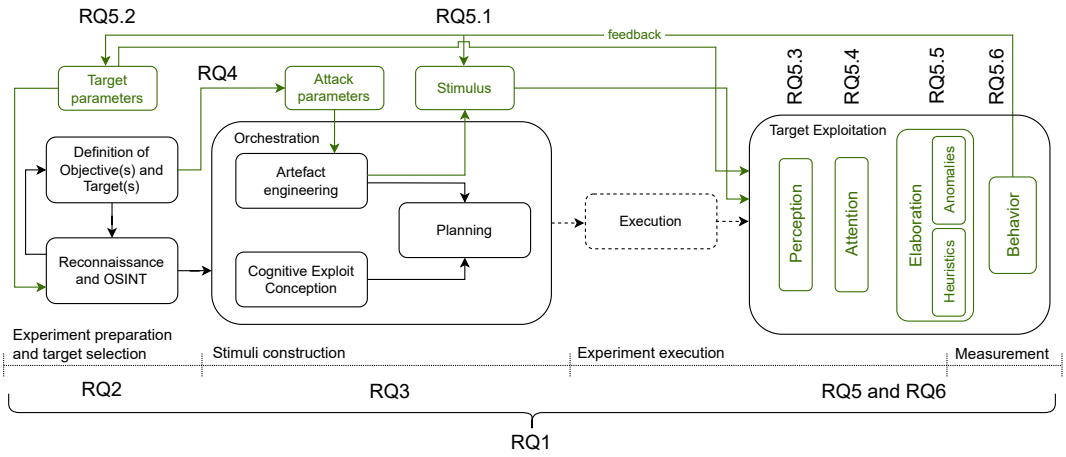

Fig. 3. Overview of the research questions.

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, artefacts and their tailoring to the subject population (reflecting 'Attack characteristics' in Table 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 [154]. In the SE context, this is particularly relevant because experiment outcomes are largely affected by the characteristics of the target population [33, 162]. The choice of the target population also reflects the subject selection and reconnaissance stages of an SE attack process (see Fig. 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. 1); thus, the artifacts used in an experiment plays 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 [81, 113]. We therefore ask:

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

The tailoring of artefacts towards the targeted population (i.e., the alignment between attack and target parameters, in terms of the cognitive framework presented in Section 2.2) is becoming an increasingly common practice in real-world attacks [79, 153] and has been shown to have a significant impact on the outcomes of SE experiments [69, 162]. The next question aims to explore the relation between target population and phishing artifacts to shed the 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 [6]. Therefore, the empirical investigation of cognitive effects is a necessary step to understand the underpinning mechanisms that bring 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 [46, 53, 199]. These represent the "SE attack surface" which encompasses the ways an attacker can deceive the target to accomplish her goals and can be characterized along the dimensions of: stimuli attributes, target characteristics (θ^p, θ^w, cognitive processes) and the contextual situation around the target (θ^s). 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 Section 2.2:

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 lead to mixed effects on behavior [95, 139] and elaboration [66]. These interactions are also affected by the level of tailoring of an artifacts [65, 81]. We therefore posit the following research question:

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

3.2 Paper collection

To cover the wide and interdisciplinary SE landscape, we choose 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 indexes the highest numbers of unique articles in computer science [38] and provides a wide coverage of venues (journals ad conference proceedings) [57], 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 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.

[1] From the best of our knowledge, the only top security venue not indexed in the Scopus database is NDSS.
[2] We did not observe any substantial differences when executing the query with more specific keywords such as `vishing`, `smishing`, etc.

Table 3. 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

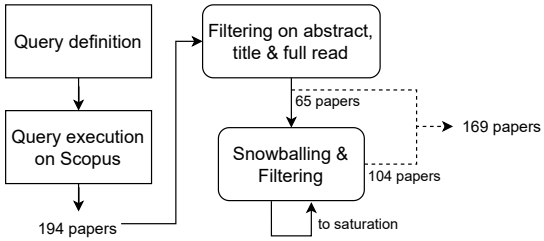

Fig. 4. Paper selection process.

To be included in the review, a paper must satisfy the criteria in Table 3. 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) [123]. A study that does *not satisfy* the criteria is an evaluation of phishing detection algorithms (not satisfying the 3rd criterion) [176] or a survey measuring the self-reported victimization and connected factors (not satisfying the 4th criterion) [39].

We applied snowballing to the papers retrieved from Scopus and meeting the criteria in Table 3 till saturation is reached, i.e., when the snowballing process yields no additional papers. 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 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 website. The summary of the whole procedure is shown in Fig. 4. The filtering step in the figure corresponds to the application of the criteria of Table 3 on title and abstract first, and on the full reading afterwards. 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. Whenever ambiguities arose in reviewing a paper (e.g., border cases), the article was analyzed and discussed between the authors till an agreement 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.

3.3 Evaluation Approach

Empirical approaches. To characterize the study design (RQ1–4), we categorize the *experiment 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 [162], 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 locations [169];

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' [113] or 'reasons for behavior' [115], in contrast to, e.g., only demographics [62].

Cognitive Features. The criteria used to answer RQ5 and RQ6 directly stem from the features of the cognitive framework presented in Section 2.2. 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 Section 2.2. 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, domain of application, artefact 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 3 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 target parameters (θ^p, θ^w, θ^s) 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 not used as treatments or control variables.

Perception. We indicate whether the study investigates any effects on perception, e.g., in terms of 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 [24]; on the other hand, an example of priming is the influencing of participants with the notion of phishing before a phishing classification task [137].

Attention. We report when the study focuses on effects related to attention, i.e., as a (in)dependent variable (`RQ5.4`). To better characterize attention, we also extracted which types of attention is engaged during the attacks, since this influences Elaboration and conditions the final behavior. We identify two possible (central) attention types as inferred from the experiment design: *exogenous* when the supplied stimuli are not part of the current activity goals of a target and *endogenous* when the target is attending the stimuli as part of their activity goals [32]. 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 [125]. For example, the use of *exogenous* attendance typically occurs in phishing susceptibility exercises at companies where the targeted employees are busy attending 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 labelled 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 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 [128, 191, 197]. 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 [20, 74]. 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 [49]. 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 experiment types employed to measure them, as this allows us to get valuable insights on 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 depended 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., an hypothesis postulating that the usage of a persuasion technique in a message (*stimulus attribute*) has some effect on *elaboration*, and controls for subjects' age (θ^p), is counted as one interaction between *stimuli attributes* and *elaboration*, one between *stimuli attributes* and θ^p and one between θ^p and *elaboration*.

4 RESULTS

In this section we present the analysis of our literature study. Results are presented by following the research questions specified in Section 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 not 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., modelling multiple covariates in a study). A comprehensive categorization of the literature sample is available in the supplementary material.

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

Fig. 5 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 experiment type; for example, 19% of

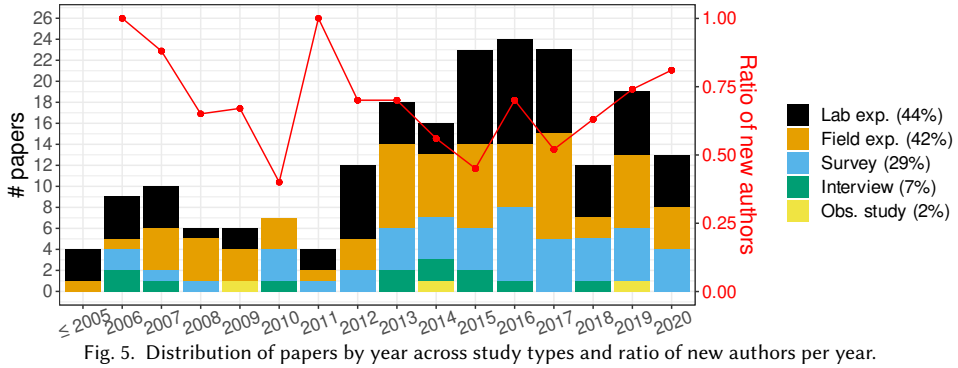

Fig. 5. Distribution of papers by year across study types and ratio of new authors per year.

all field and lab experiments are complemented by a survey to capture additional variables and relevant factors needed for the testing of one ore more hypotheses, such as participants' susceptibility awareness [113] or their reasoning patterns [115]. 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 [73]. Another study estimated actual and future clicks on links in real phishing campaigns as a function of persuasion techniques employed in the emails [175]. 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 relative low number of such studies in the extant literature.

Our analysis reveals that the number of unique authors per year follows a distribution similar to the one reported for papers in Fig. 5. The ratio of new authors per year (red line in Fig. 5) shows that, during the whole period, at least half of the authors each year were new, bringing the total number of unique authors as per 2020 to 371. Among authors there seems to be some "specialization" in specific empirical methodologies: anti-phishing awareness and training methods carried out with mix of lab and field experiments (e.g., [53, 72, 104, 158]); attack susceptibility via email and SNS, effects of heuristic processing (e.g., [77, 184, 185, 188]) and persuasion techniques (e.g., [116, 198, 200]) which were mostly investigated with field experiments; and effects of priming [137, 139] and training efficacy (e.g., [165, 187]) tested in lab experiments. We observe an overall decrease in activity in terms of output volume around 2018. Untangling the reasons for such dynamics is out of scope of this review. Nevertheless, a possibility is that there is a relative 'saturation' of research questions that can be tested with techniques already familiar to researchers. In that regard, insights from this literature survey aim at opening the field to new research directions, expanding the scope of empirical research in this area. A discussion is provided in Section 5. An analysis of the sample size employed in the extant literature is reported in the supplementary material.

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

Fig. 6 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 [108, 165]. The second most frequent target population are 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

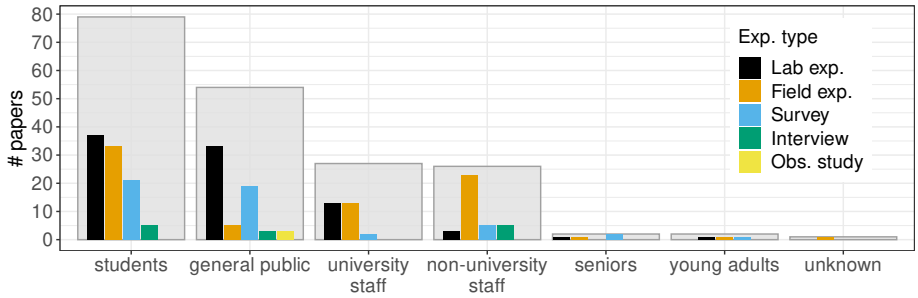

Fig. 6. Distribution of papers by targeted population across study types.

users [24, 182] or eBay users [89]. 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 [125] and a field experiment with older vs. young adults [113]. 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 [202]).

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 [19], suggesting that academic studies and experiments may be overall of only limited relevance for 'real world' attacks. For example, findings in [33] and [102] 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 is for the great 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 type 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 organizations employees, and across domains, remains open. Further, highly vulnerable categories such as senior citizens and youngsters remain widely understudied.

4.3 RQ3: What types of artefacts have been considered for the delivery of social engineering attacks in empirical SE literature?

Fig. 7 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

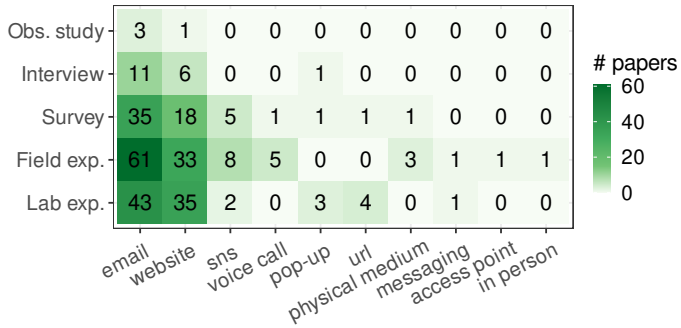

Fig. 7. Distribution of papers with respect to stimuli and study types.

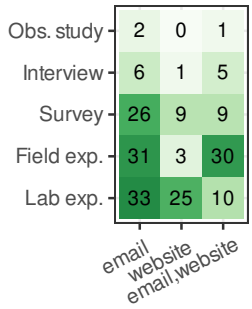

Fig. 8. Zoom-in on combinations of email and website stimuli types.

prevalent. Other stimuli include Wi-Fi access points and instant messages [98], pop-ups [125], physical media (brochures, mail, affixed QR codes) [95, 179, 195] and in person deception [30]. 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 under-represented despite these being increasingly often reported in the wild [19]; examples are deception over voice [22] or Stuxnet-like attacks [106], and more in general multi-step-multi-media attacks, such as reverse SE on LinkedIn [6] or lateral movement attacks in organizations [41]. Studies reproducing these attack scenarios are largely not yet reported in the literature.

As email and websites are tightly linked and generally part of the same attack procedure (e.g., email with a URL linking to a counterfait login interface), Fig. 8 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) [46, 199]. On the other hand, this is generally not the case in reality [63]. Therefore, these studies may not accurately capture real victimization rates. For example, tech-savvy 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 [62]. 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[3]). We observe that the majority of these are field experiments, for example simulating phishing campaigns with a website asking some information. Interestingly, there are only few lab experiments in our sample that reproduce two-step SE attacks, e.g. [8, 108, 201]; the sheer majority of lab experiments feed participants with one stimulus at the 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 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 into this direction [56, 135, 163]. From Fig. 8, we can observe only three field experiments that use websites in combination with other stimuli types (field

[3]Among the few works that did combine email,website and other stimuli, Workman et al. [194–196] utilized a combination of email, website, voice call (and mail) but without providing details on the specific behavior(s) being measured.

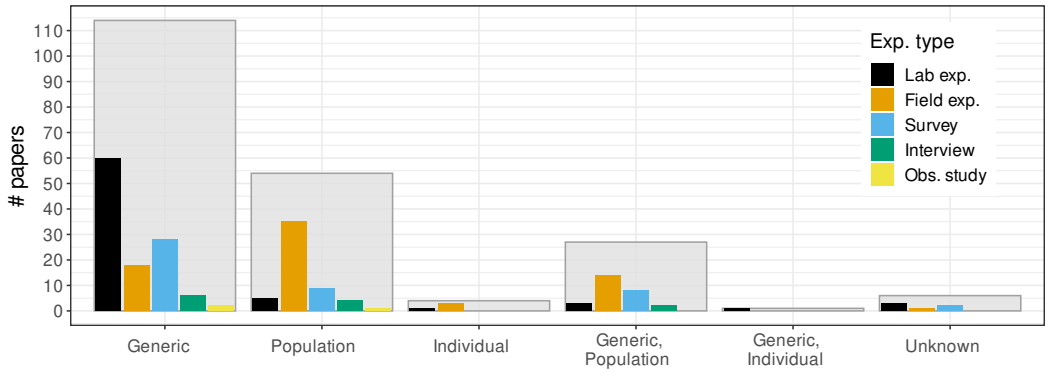

Fig. 10. Distribution of papers by attack targetization across study types. G=generic, P=population, I=individual, U=unknown

experiments using only websites are absent due to the practical limitation of delivering websites to subjects without other media); these studies use access points [98], SNS [160] and QR codes [179].

Fig. 9 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 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 [160]. This emphasizes 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 [32], can help defining a coherent and consistent account of possible interactions between multi-stage stimuli, and help identify relevant target population and methodological choices to employ for their investigation.

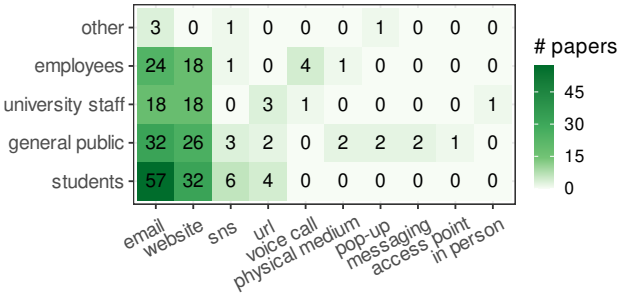

Fig. 9. Distribution of papers with respect to stimuli types and targeted populations.

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

Fig. 10 shows the distribution of attack targetization over experiment 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 [170] or of training against spear-phishing attacks [34]. By contrast, recent attacks already show signs of automated tailoring, such as automatic detection of the affiliated company based on the domain in the user's email address and integration of that company's logo into a fraudulent webpage [166]. We expect this type of attacks to be observed more often in the wild due to the possibility of automatically scraping data of potential targets across OSINT sources and data leaks [6], which can be semi-automatically exploited using specialized toolkits for tailored phishing [146]. This positions the current state

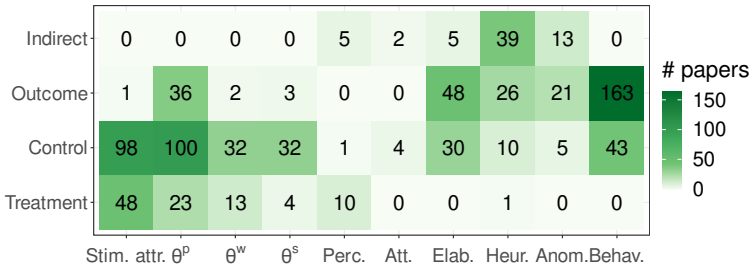

Fig. 11. Distribution of framework features in the experiments and their employment as treatment, control, and outcome variables; variables only included indirectly in a study design are reported as 'indirect' in the figure.

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. 10, 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. [81] 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 [9, 10, 194, 196], suggesting that targetization is an important variable to account for in SE studies. This is contrasting 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 [69]. To address this, future studies could benefit from a consistent accounting of the adaptation degree between stimuli, targeted population and, where possible, the target context. With respect to the framework in Fig. 2, this translates to determining the degree of matching between target and attack parameters and account for such a degree during artifact construction.

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

Fig. 11 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 (θ^p) 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, only 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 (θ^p, θ^w, θ^s) are for the largest part used as controls. However, θ^p and θ^w have also been instrumented as treatments, often in the form of anti-phishing training/awareness in either personal or work-related contexts. θ^p 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. [11, 126]. θ^s have been almost exclusively used as control variables; when used as treatments, they were employed in the form of

incentives provided to the participants [110, 127, 159, 205]. 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., [24, 36]) or indirectly trigger the activation of generic vs. specific percepts with highly contextualized stimuli (e.g., [65, 81]). We find that attention is the least investigated feature in empirical SE: two studies indirectly manipulated attention [186, 188] and other four studies used attention as a control in relation to the outcome variable, e.g. [125, 192].

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., [126, 179]). Heuristics and Anomalies features are predominantly studied as indirect effects of cognitive biases and anomalies [20], and as outcome variables in terms of heuristics and trust indicators [46, 186].

The overview of the status of empirical SE research reported above indicates that there might be certain 'boundaries' with respect to what is being measured and what is possible to measure with the available techniques. For example, investigating effects of stimuli attributes and target parameters (θ) 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 [147].

4.5.1 *RQ5.1. What stimuli attributes have been investigated?*

Fig. 12 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 [2, 28, 30, 194, 196] or social network interactions [180, 182, 184] (category 'other' in Fig. 12). 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 [6].

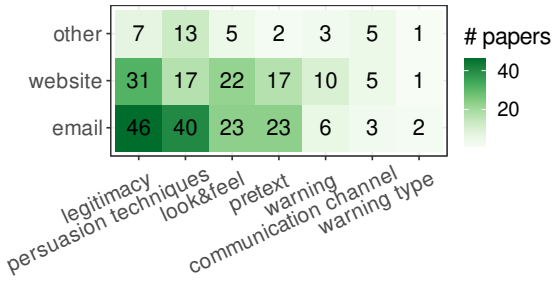

Fig. 12. Distribution of papers with respect to stimuli attributes and stimuli types.

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 stimuli legitimacy [99]. This method is a popular and easy way to estimate the susceptibility to SE attacks of a given population (cf. Section 4.5.6). 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 to investigate deeper given their widespread and the great potential for misuse [164]. There is thus a lack of studies exploring the effects on elaboration and behavior of otherwise commonly investigated stimuli attributes (excluding persuasion techniques)

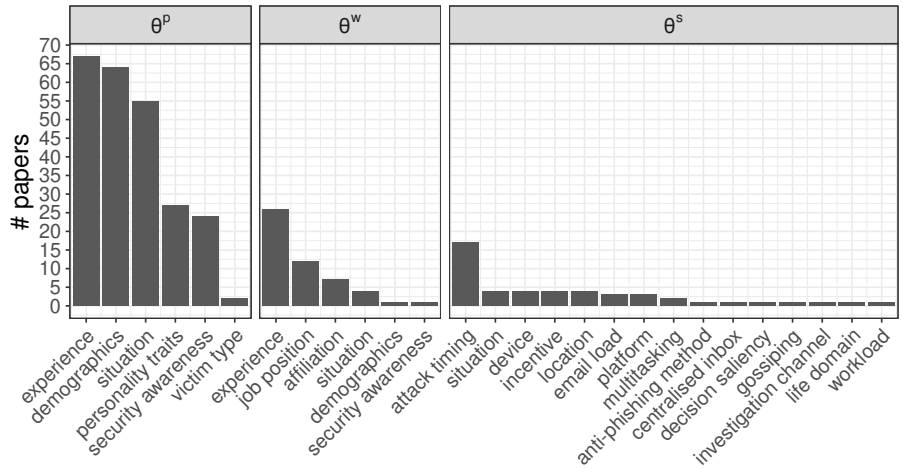

Fig. 13. Distribution of papers by parameter type and category.

on SNS, instant messages or voice calls. As mentioned in Section 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, 50, 55, 203]). Such studies often involved testing warning messages (as showed in Fig. 12) aimed at preventing user deception. However, the *type* of warning message, 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 [51, 187, 201]. 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 un-intrusive warnings and interventions (e.g., *nudges*) has been recently highlighted as a valuable opportunity to improve current defense techniques [64].

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

Fig. 13 shows the distribution of the studied target parameters. It is immediate to observe that the sheer majority of investigated parameters fall into personal subject parameters (θ^p). 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 [67] or self-efficacy [204] 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 θ^s which regard the *contextual* dimension of the circumstances of the experiment, such as the environment [161], timing [52] or the concurrent activities of the participants [186]. Other personality traits represent long-term factors, such as propensity to trust or the *Big Five Inventory* (BFI) personality measures [94]. A number of personal subject parameters show consistent negative effects on SE susceptibility, such as experience [199] and knowledge [76, 188], whereas others show mixed or no effects (e.g., age [137, 194], gender [137]). Subject parameters such as curiosity and commitment are often reported to have a positive effect (i.e., they *increase* attack susceptibility) [123, 194].

Professional target parameters (θ^w) mostly consider subjects' professional experience (15%) such as years of service and security training in the working environment [96, 190]. Job position (e.g., student, professor, management, support staff) is oftentimes used as a proxy for familiarity with the overall organization context [33]; however, this may introduce

errors for newly hired professionals in senior positions [29]. Only one study in our sample evaluated effects of subject parameters across different organizations (yielding mixed outcomes) [33].

The most common setting parameter (θ^s) 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 [183, 184] and one lab experiment [121]; 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 [121], 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 [69].

Overall, our first most important observation is the net tendency of the state-of-the-art to favour θ^p and θ^w characteristics in their investigations. Whereas individual personal and work-related factors are undoubtedly important in affecting the outcome of SE attacks, setting parameters (θ^s) are largely dismissed as of secondary importance both in terms of quantity 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 [69]. For example, only four studies considered smartphone usage (grouped under `device` in Fig. 13), which is surprising considering the popularity of mobile phones and the usability constraints they introduce both at the interface level [183] and at the contextual level [159, 183] (e.g. multitasking while on the go).

The lack of studies considering setting-related factors may be partly due to the inherent difficulty to control for such variables. Nevertheless, some studies were able to measure some aspects of the target's context, such as workload [91] or email load [191], or the life domains the targets are sensible to [113]. 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 [117]. 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 [69].

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

Perception deals with the translation of stimuli into percepts. This process can bring to a general or a more specific set of percepts being loaded in a subject's working memory, depending on how well the stimulus is aligned with the target's context [32]. Among the possible effects is the conditioning of perception with *pre-attack* and *priming* operations prior the delivery of deceptive stimuli (cf. Section 2.2). Only a few papers in our collection study such effects either for defense [36, 66, 68, 87, 93, 95, 137, 139, 141, 142] or attack [24]. From these studies, the effect of priming is unclear. For example, Benenson et al. [24] do not find significant effects of priming on attack success (i.e., sending SNS friend requests before actual attack). Many studies on priming in defensive scenarios [36, 66, 137, 139, 141, 142] do find that priming has a significant positive effect, while other studies [36, 68, 87, 95] 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 [137]. 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 [48] and subliminal triggers of affective reactions [193].

Only a few studies investigated the specificity of target-related information and contextualization in phishing attacks [65, 81, 84, 99]. 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 [34, 170]. 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 [152] may be used to, for example, 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 focusses on information-rich media, such as in person or verbal communication, the SE literature has focussed mainly on written communication forms, i.e., emails and websites. This suggests that effects on perception may be partiularly 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 attacker's success over voice calls [172], as also exemplified by recent vishing attacks in political cases [22] and recent mass social security scams involving expats [54]. Similarly, the involved perceptual mechanisms in written attacks still have the potential to make scams more convincing, such as friends recommendations on Facebook [85] or the specificity of tailored scenarios [6]. We argue that this represents an opportunity to define a new research line to test and address new, unconventional forms of attacks involving perception.

4.5.4 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. Section 2.2). From Fig. 11, we can observe that only six works studied the effects on attention: five studies [125, 186, 188, 192, 198] investigated the effect of attention type (endogenous, exogenous) and one measured the level of (endogenous) attention [149]. Wang et al. [188] and Wright et al. [198] 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. [125] 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 [149] indicate a strong positive correlation between the ability to exercise sustained attention (closely related to endogenous attention [155]) 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. 14 presents a breakdown of attention types with respect to experiment type. Laboratory experiments (and surveys) almost always implies the use of endogenous attention, while field experiments employ exogenous attention. Studies marked as '*ex,end*' involve both types (e.g., [2, 82, 91]) or control for attention type [125, 192].

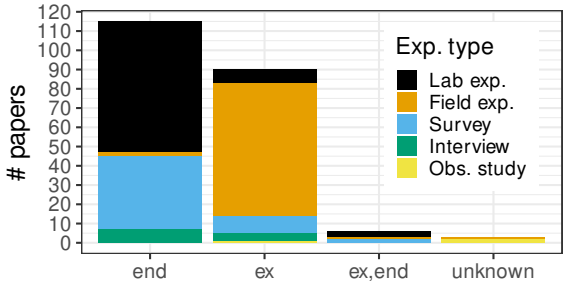

Fig. 14. Distribution of papers by attention type across study types.

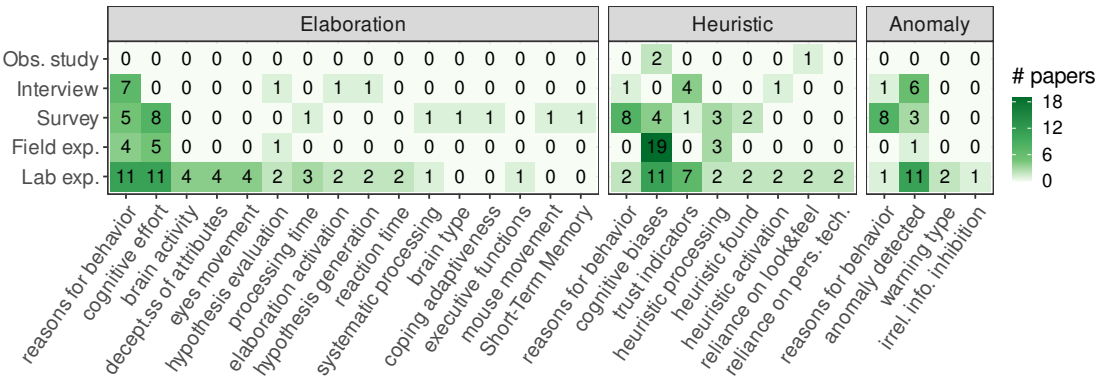

Fig. 15. Distribution of papers by features related to Elaboration, Heuristic and Anomaly w.r.t. study types.

Overall, these observations indicate that attention can play a decisive role on 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 [125]) and related to, e.g., matching of target parameters (as in [183]). Techniques to manage attention can be adopted from other fields, such as parallel recognition tasks from cognitive sciences [111]. 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 [64].

4.5.5 *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. 15. 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. [138] included open questions about participants' reasoning for decision making to develop a framework on user intention and actual behavior; Nicholson et al. [131] tested anomaly detection with saliency nudges as treatments in an online lab task. Similarly, but in a more elaborate lab setup, Hale et al. [74] explored heuristic activation and anomaly detection. Field experiments have also been adopted to study elaboration features, especially concerning *Heuristics*: Williams et al. [191] 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. [37] 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

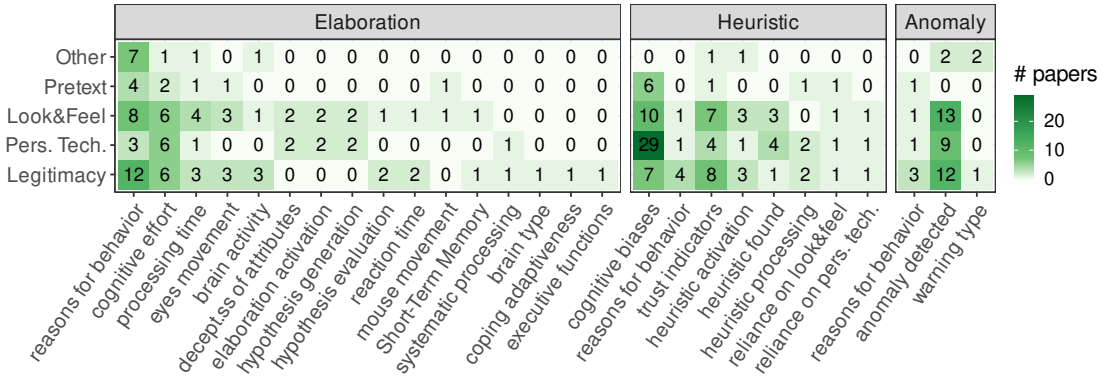

Fig. 16. Distribution of papers by features related to Elaboration, Heuristic and Anomaly w.r.t. stimuli attributes.

such as heuristics usage, e.g. [185, 186], trust indicators or anomaly detection, e.g. [90, 143]. 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) [133]. Nonetheless, more direct measurements have also been adopted in certain cases. For instance, cognitive effort has been measured as a function of time [137] or by means of eye-tracking devices [128], which are employed to identify visual focus areas during the elaboration [120]. Additionally, several attempts to directly measure brain activity have been carried out with, e.g., fMRI during phishing classification tasks [129, 130, 174], 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 [130]. However, performing such measurements is challenging as effects remain difficult to isolate [188].

Fig. 16 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 uncertainty, several studies implemented additional measurements; for example, Parsons et al. [136] 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. [184] 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 [112] 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 [20] 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 [65]. In accordance with previous literature [162], 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 [69, 113].

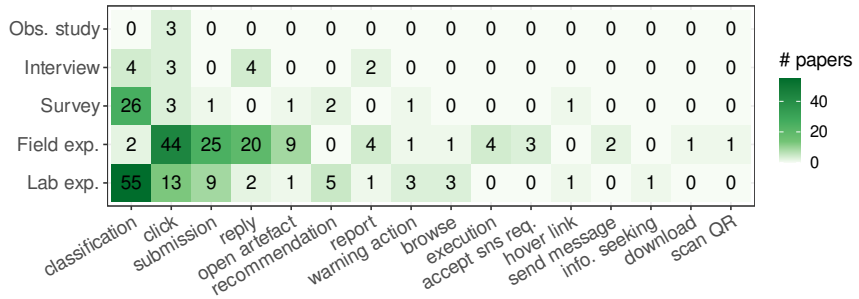

Fig. 17. Distribution of papers by behavior w.r.t. study types.

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) [133], or develop methods able to measure features when the processing phase is still 'hot', that is, immediately after a certain action is performed [64] (e.g., instant feedback [163]). The present state of knowledge in social cognition provides a plethora of implicit measurement methods and research evidence as summarized in [133], with applications spanning from organizational [173] to consumer research [151]. The applicability and adaptation of such techniques to the SE domain remain, however, an open question.

4.5.6 *RQ5.6. What types of behavior have been investigated?*

Fig. 17 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., [37, 97, 188]), 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 [101]), 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 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' [43, 60, 177]. Yet, the assumption that a click corresponds to a security impact is oftentimes (explicitly) made: "*clicking [...] deploys malware and opens virtual backdoors*" [181], "*the click of an email link can take users to a fake site requesting login information*" [84] or "*expose the organization to a network of hackers*" [91]. None of these works, however, modeled or measured the actual compromise, information submission or exploitation by means of, e.g., submission forms or executables [63, 102]. Whereas relevant, the implicitly assumed threat model diverges significantly from that of a 'regular' attacker (see, e.g., [26]), 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 [92, 163], other two PDF files with links [183, 185] and one with an HTML attachment [51], whereas we find no study employing MS Office documents, despite their importance as delivery vector of malware [140].

A number of studies do distinguish link clicks from submission of information (e.g., [81, 102]), submissions only (e.g., [33, 76]) or consider the opening of artefacts and executables (e.g., attachments [185] or downloaded files [81]) as measures of 'success'. Generally, experiments considering two-stage scenarios (e.g., phishing email and a subsequent

landing webpage, see Section 4.3) report lower success rates than one-stage studies [162]. 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 [37, 51, 70, 197], despite reports being a valuable prevention and mitigation method [31].

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 behavior, indicators of trust in stimuli and identified anomalies (e.g., [37, 53, 199]). Interestingly, two studies report remarkable differences between *intention* to click and *actual* clicks, finding rates in the latter higher than in the former [82, 113]. However, surveys and interviews often do not consider other relevant behaviors, such as credential submissions and opening/executing artefacts. Neither do lab experiments, except a small user study with attachment-like artefacts [163]. Nevertheless, such experiments can be valuable instruments to capture user perception or decision making, e.g., trust certain file types, enable macros in documents files 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 [185]. 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 for guide the user in the right decision for such cases.

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

Fig. 18 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 an 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 θ^p, as well as personal parameters θ^p and work parameters θ^w, 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 [4, 65], or with job position and type of anti-phishing training [93, 124]. Similarly, attributes and θ^p 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* [181, 191], as also discussed in Section 4.5.5).

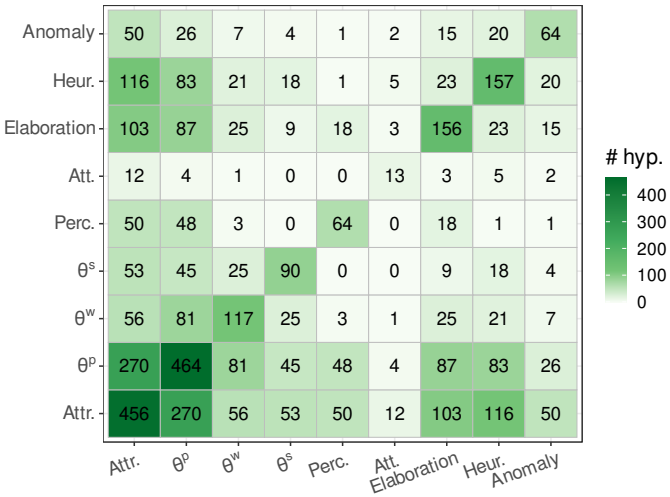

Fig. 18. Studied interactions between SE cognitive features.

Studies investigating the interactions between perception and other cognitive features are, in general, rare (cf. Section 4.5.3). The only interactions that have been studied are with the stimulus attributes (e.g., [95]), θ^p (e.g., [66, 137]) and to a lesser degree with *Elaboration* (e.g., [139]) and θ^w (e.g., [93]). 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, of the attack delivered through the LinkedIn platform described in [6].

θ^s is 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 θ^s effects on *Elaboration* (e.g., [86, 127, 159]). Similarly, we find large gaps between cognitive processing at *Elaboration* and attention levels (e.g., [35, 149]), and little work explicitly studying the relation between anomalies and heuristics with *Elaboration* dynamics (e.g, [76, 185]). Yet, insights on interaction effects with *Elaboration* and other features can hold valuable implications towards 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) [76]. 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 [22] where the deceiver deceives his target into revealing information by means of authoritative impersonation (spoofing caller ID) and overloading the target's attention with several contextual details. These examples underline the relevance of gaps in the map of Fig. 18, which clearly shows that the literature has been focusing on a rather narrow research space.

5 DISCUSSION

In the previous section, we have analyzed the collected studies on the various dimensions represented by our research questions, as outlined in Fig. 3. In this section we summarize our findings and identify gaps in the literature and promising directions for future work, thus answering our main research question.

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 a remarkable precision in simulating the archetype of SE attacks: a phishing email with a malicious link. However, there are many more scenarios that is worth investigating deeply, which are largely ignored by the literature (cf. Section 4.3). Moreover, several simulated phishing campaigns and classification tasks do not reflect the current threat landscape (cf. Section 4.5.6). 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 [15]) or opening attachment-like artefacts. Lab experiments should accurately reproduce the complete attack process, e.g. click then submit, and allow for interactive interfaces (such as hover on links, reactive forms, etc.) over static screenshots of emails and websites. Further, modern-day attacks feature a diversification of utilized media (cf. Section 4.3) and of the modality of their employment (cf. Section 4.5.4). 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 [72], social networks and email [6] or QR codes [42, 179]. 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 conduced with participants drawn from university pools (cf. Section 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 [19]. 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 [33, 69, 105]. 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 towards 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 regard *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 4.5.2, profession-related (θ^w) and setting-related (θ^s) 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 [69, 162], 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 4.5.3, 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 4.5.4), 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 4.5.5). 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 [76], it remains unclear under which circumstance it is triggered during processing. Similarly, the power of 'anchoring' effects such as cognitive biases are 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 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 to expand previous approaches to new or adapted techniques from cognitive science and social psychology (e.g., [111, 133]) with the aim of capturing and analyzing 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 4.2 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). 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 4.5). This saturation of what can be measured or tested does not help filling 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 4.5.2) 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 [3, 206]. 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 [24] and [126] 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 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 [69], as also thoroughly discussed in Section 4.5.2. 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 [65, 72, 88]. 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 severs than they were to email with links to internal severs [27], and where attempts to individualize adaptations (e.g., saluting the recipient by name [89] or congruently to expectations of participants [132]) were no more successful than generic emails [162]. 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 [32], 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 4.5.6. Clicks or wrong classification of stimuli may result in a successful outcome for an attacker, and studies employing such measurements provide valuable results [141, 157]. 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 [91], behavior sciences [84]) 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. Organizations use the results of simulations, i.e. click rates, to draw conclusions on their information security posture and for policy making [100, 150]. When not applied carefully (e.g., in the design of embedded phishing training), this can lead to adverse side effects, such as a false sense of security, making employees even more vulnerable to phishing [105].

Relevant factors often not controlled for. As we have seen in Section 4.5, some factors are especially difficult to control or measure, namely setting parameters (θ^s) as well as cognitive features such as *Perception*, *Attention* and *Elaboration*. 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 [13]. 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 [117], the role of priming [48, 193] and memory [156] in subjects' perception or how their elaboration can be influenced by attention [111], heuristics [23, 71] and anomalies [83]. Such an interdisciplinary approach to solve problems in information security is well exemplified by the results obtained in usable security studies where concepts and techniques, such as behavior change theories [118] and digital 'nudges' [168], have been successfully employed in experiments to evaluate, for example, warning [112, 131] or training efficacy [102, 161], as also discussed in [64, 144].

5.1 Threats to validity

Internal threats. The cognitive framework adopted for our analysis [32] 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 [122].
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 a limitation by performing a reverse snowballing till saturation was reached, and by running more specific queries on Scopus which showed no substantial difference in results. 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. 16. 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 consensus was reached. More in general, this a common problem in similar articles, where the absence of an established framework for classifying SE experiments poses such limitations [162]. To this end, we distilled our criteria from a well-established cognitive framework specifically designed for the analysis of SE attacks [32].

6 CONCLUSION

This work provided a systematic review of the state of empirical SE research on cognition with the goal of advancing the body of knowledge in the Social Engineering 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 multidisciplinary landscape of empirical SE research along the dimensions of experiment design and human cognition. By relating the findings of the analysis with the dynamics of real attacks and extant SE research, we identified relevant insights and promising directions for future work.

REFERENCES

[1] A. Abbasi, F. Mariam Zahedi, and Y. Chen. 2016. Phishing susceptibility: The good, the bad, and the ugly. In *International Conference on Intelligence and Security Informatics*. IEEE, 169–174.

[2] M. Aburrous, M. Hossain, K. Dahal, and F. Thabtah. 2010. Experimental Case Studies for Investigating E-Banking Phishing Techniques and Attack Strategies. *Cognitive Computation* 2, 3 (2010), 242–253.

[3] D. Akhawe and A. Felt. 2013. Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness. In *USENIX Security*. 257–272.

[4] A. Algarni, Y. Xu, and T. Chan. 2017. An empirical study on the susceptibility to social engineering in social networking sites: The case of Facebook. *Eur J Inf Syst* 26, 6 (2017), 661–687.

[5] L. Allodi. 2017. Economic Factors of Vulnerability Trade and Exploitation. In *CCS*. ACM, 1483–1499.

[6] L. Allodi, T. Chotza, E. Panina, and N. Zannone. 2020. The Need for New Antiphishing Measures Against Spear-Phishing Attacks. *IEEE Security Privacy* 18, 2 (2020), 23–34.

[7] L. Allodi and F. Massacci. 2014. Comparing Vulnerability Severity and Exploits Using Case-Control Studies. *TISSEC* 17, 1 (2014), 1:1–1:20.

[8] A. Alnajim and M. Munro. 2009. An Anti-Phishing Approach that Uses Training Intervention for Phishing Websites Detection. In *International Conference on Information Technology: New Generations*. 405–410.

[9] I. Alseadoon, T. Chan, E. Foo, and J.G. Nieto. 2012. Who is more susceptible to phishing emails?: A Saudi Arabian study. In *Australasian Conference on Information Systems*.

[10] I.M. Alseadoon, M.F.I. Othman, E. Foo, and T. Chan. 2013. Typology of phishing email victims based on their behavioural response. In *Americas Conference on Information Systems*, Vol. 5. 3716–3724.

[11] I. Alseadoon, M. F. I. Othman, and T. Chan. 2015. What Is the Influence of Users' Characteristics on Their Ability to Detect Phishing Emails?. In *Advanced Computer and Communication Engineering Technology (LNEE)*. Springer, 949–962.

[12] Amnesty International. 2019. *Evolving Phishing Attacks Targeting Journalists and Human Rights Defenders from the Middle-East and North Africa*. Retrieved March 22, 2022 from https://edu.nl/7t9wd

[13] B.B. Anderson, A. Vance, C.B. Kirwan, D. Eargle, and J.L. Jenkins. 2016. How users perceive and respond to security messages: A NeuroIS research agenda and empirical study. *Eur J Inf Syst* 25, 4 (2016), 364–390.

[14] J. Anderson. 2000. *Cognitive psychology and its implications*. Worth publishers.

[15] Arstecnica. 2022. *Lapsus$ and SolarWinds hackers both use the same old trick to bypass MFA*. Retrieved March 29, 2022 from https://edu.nl/qdwj8

[16] B.J. Baars. 2002. The conscious access hypothesis: Origins and recent evidence. *Trends in Cognitive Sciences* 6, 1 (2002), 47–52.

[17] B.J. Baars and S. Franklin. 2003. How conscious experience and working memory interact. *Trends in Cognitive Sciences* 7, 4 (2003), 166–172.

[18] A. Baddeley. 2000. The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences* 4, 11 (2000), 417–423.

[19] Barracuda. 2021. *Spear Phishing: Top Threats and Trends*. Technical Report.

[20] M. Baryshevtsev and J. McGlynn. 2020. Persuasive Appeals Predict Credibility Judgments of Phishing Messages. *Cyberpsychol Behav Soc Netw* 23, 5 (2020), 297–302.

[21] BBC News. 2018. *Twelve Russians charged with US 2016 election hack*. Retrieved March 22, 2022 from https://edu.nl/98mux

[22] Bellingcat. 2020. *FSB Team of Chemical Weapon Experts Implicated in Alexey Navalny Novichok Poisoning*. Retrieved March 22, 2022 from https://edu.nl/vaxy9

[23] S. Bellur and S. Sundar. 2014. How Can We Tell When a Heuristic Has Been Used? Design and Analysis Strategies for Capturing the Operation of Heuristics. *Commun Methods Meas* 8, 2 (2014), 116–137.

[24] Z. Benenson, F. Gassmann, and R. Landwirth. 2017. Unpacking spear phishing susceptibility. In *Fin. Crypto. & Data Sec. (LNCS)*. Springer, 610–627.

[25] M. Bitaab, H. Cho, A. Oest, P. Zhang, Z. Sun, R. Pourmohamad, D. Kim, T. Bao, R. Wang, Y. Shoshitaishvili, A. Doupé, and G. Ahn. 2020. Scam Pandemic: How Attackers Exploit Public Fear through Phishing. In *Symposium on Electronic Crime Research*. 1–10.

[26] S. Blond, A. Uritesc, C. Gilbert, Z. Chua, P. Saxena, and E. Kirda. 2014. A Look at Targeted Attacks Through the Lense of an NGO. In *USENIX Security Symposium*. USENIX Association, 543–558.

[27] B. Bowen, R. Devarajan, and S. Stolfo. 2011. Measuring the human factor of cyber security. In *Int. Conf. on Tech. for Homeland Sec.* IEEE, 230–235.

[28] J. Bullee, L. Montoya, M. Junger, and P. Hartel. 2016. Telephone-based social engineering attacks: An experiment testing the success and time decay of an intervention. In *Inaugural Singapore Cyber Security R&D Conference*. IOS Press, 107–114.

[29] J. Bullee, L. Montoya, M. Junger, and P. Hartel. 2017. Spear phishing in organisations explained. *Inf. Comput. Secur.* 25, 5 (2017), 593–613.

[30] J. Bullée, L. Montoya, W. Pieters, M. Junger, and P. Hartel. 2015. The persuasion and security awareness experiment: reducing the success of social engineering attacks. *J. Exp. Criminol.* 11, 1 (2015), 97–115.

[31] P. Burda, L. Allodi, and N. Zannone. 2020. Don't Forget the Human: a Crowdsourced Approach to Automate Response and Containment Against Spear Phishing Attacks. In *EuroS&P Workshops*. IEEE, 471–476.

[32] P. Burda, L. Allodi, and N. Zannone. 2021. Dissecting Social Engineering Attacks Through the Lenses of Cognition. In *2021 IEEE European Symposium on Security and Privacy Workshops (EuroS PW)*. 149–160.

[33] P. Burda, T. Chotza, L. Allodi, and N. Zannone. 2020. Testing the Effectiveness of Tailored Phishing Techniques in Industry and Academia: A Field Experiment. In *International Conference on Availability, Reliability and Security*. ACM.

[34] A.J. Burns, M.E. Johnson, and D.D. Caputo. 2019. Spear phishing in a barrel: Insights from a targeted phishing campaign. *J. Organ. Comput. Electron. Commer.* 29, 1 (2019), 24–39.

[35] M. Butavicius, K. Parsons, M. Pattinson, and A. McCormac. 2015. Breaching the human firewall: Social engineering in phishing and spear-phishing emails. In *Australasian Conference on Information Systems*.

[36] C. Canfield, B. Fischhoff, and A. Davis. 2016. Quantifying Phishing Susceptibility for Detection and Behavior Decisions. *Human Factors* 58, 8 (2016), 1158–1172.

[37] D.D. Caputo, S.L. Pfleeger, J.D. Freeman, and M.E. Johnson. 2014. Going spear phishing: Exploring embedded training and awareness. *IEEE Secur Priv* 12, 1 (2014), 28–38.

[38] A. Cavacini. 2015. What is the best database for computer science journal articles? *Scientometrics* 102, 3 (2015), 2059–2071.

[39] H. Chen, C.E. Beaudoin, and T. Hong. 2017. Securing online privacy: An empirical test on Internet scam victimization, online privacy concerns, and privacy protection behaviors. *Comput. Hum. Behav.* 70 (2017), 291–302.

[40] R. Cialdini. 2016. *Pre-suasion: A revolutionary way to influence and persuade.* Simon and Schuster.

[41] A. Cidon, L. Gavish, I. Bleier, N. Korshun, M. Schweighauser, and A. Tsitkin. 2019. High Precision Detection of Business Email Compromise. In *USENIX Security Symposium*. USENIX Association, 1291–1307.

[42] City of Austin. 2022. *Fraudulent QR codes found on Austin parking pay stations | AustinTexas.gov.* Retrieved March 22, 2022 from https://edu.nl/xpev8

[43] Cofense. 2020. *COFENSE Q3 Phishing Review.* Technical Report. 13 pages.

[44] A. Darwish, A. Zarka, and F. Aloul. 2012. Towards understanding phishing victims' profile. In *ICCSII*. IEEE, 1–5.

[45] S. Dehaene and L. Naccache. 2001. Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition* 79, 1-2 (2001), 1–37.

[46] R. Dhamija, J. D. Tygar, and M. Hearst. 2006. Why phishing works. In *SIGCHI Conference on Human Factors in Computing Systems*. ACM, 581–590.

[47] T. Dijk. 2008. Context and cognition. In *Discourse and Context: A Sociocognitive Approach*. Cambridge University Press, 56–110.

[48] A Dijksterhuis and J. Bargh. 2001. The perception–behavior expressway: Automatic effects of social perception on social behavior. In *Advances in experimental social psychology*. Vol. 33. Academic Press, 1–40.

[49] A. Dimoka, P. Pavlou, and F. Davis. 2011. Research Commentary—NeuroIS: The Potential of Cognitive Neuroscience for Information Systems Research. *Inf. Syst. Res.* 22, 4 (2011), 687–702.

[50] David G. Dobolyi, Ahmed Abbasi, F. Mariam Zahedi, and Anthony Vance. 2020. An Ordinal Approach to Modeling and Visualizing Phishing Susceptibility. In *International Conference on Intelligence and Security Informatics*. IEEE, 1–6.

[51] R. Dodge, C. Carver, and A. Ferguson. 2007. Phishing for user security awareness. *Computers & Security* 26, 1 (2007), 73–80.

[52] R. Dodge, K. Coronges, and E. Rovira. 2012. Empirical Benefits of Training to Phishing Susceptibility. In *Inf. Sec. & Priv. Research*. Springer, 457–464.

[53] J. Downs, M. Holbrook, and L. Cranor. 2006. Decision strategies and susceptibility to phishing. In *SOUP*. ACM, 79–90.

[54] DutchReview. 2021. *You're under arrest: thousands of Dutchies targeted by phishing calls.* Retrieved March 22, 2022 from https://edu.nl/gwjpg

[55] S. Egelman, L. Cranor, and J. Hong. 2008. You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In *SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1065–1074.

[56] S. Egelman and S. Schechter. 2013. The Importance of Being Earnest [In Security Warnings]. In *Fin. Crypto. and Data Sec. (LNCS)*. Springer, 52–59.

[57] Elsevier. 2020. *Scopus Content Coverage Guide.* Retrieved March 22, 2022 from https://edu.nl/wxmgh

[58] Endgadget. 2019. *Evidence mounts that Russian hackers are trying to disrupt the EU elections.* Retrieved March 22, 2022 from https://edu.nl/a69rb

[59] J. Evans. 2003. In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences* 7, 10 (2003), 454–459.

[60] F-Secure. 2021. *Attack Landscape Update.* Technical Report. 29 pages.

[61] A. Ferreira, L. Coventry, and G. Lenzini. 2015. Principles of Persuasion in Social Engineering and Their Use in Phishing. In *Human Aspects of Information Security, Privacy, and Trust (LNCS)*. Springer, 36–47.

[62] R. Flores, H. Holm, M. Nohlberg, and M. Ekstedt. 2015. Investigating personal determinants of phishing and the effect of national culture. *Inf. Comput. Secur.* 23, 2 (2015), 178–199.

[63] W.R. Flores, H. Holm, Gu. Svensson, and G. Ericsson. 2013. Using phishing experiments and scenario-based surveys to understand security behaviours in practice. In *European Information Security Multi-Conference*. 79–90.

[64] A. Franz, V. Zimmermann, G. Albrecht, K. Hartwig, C. Reuter, A.r Benlian, and J. Vogt. 2021. SoK: Still Plenty of Phish in the Sea — A Taxonomy of User-Oriented Phishing Interventions and Avenues for Future Research. In *SOUPS*. USENIX Association, 339–358.

[65] S. Goel, K. Williams, and E. Dincelli. 2017. Got phished? Internet security and human vulnerability. *J. Assoc. Inf. Syst.* 18, 1 (2017), 22–44.

[66] S. Grazioli. 2004. 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.

[67] S. Grazioli and S.L. Jarvenpaa. 2000. Perils of Internet fraud: an empirical investigation of deception and trust with experienced Internet consumers. *IEEE Trans. Syst. Man Cybern.* 30, 4 (2000), 395–410.

[68] S. Grazioli and A. Wang. 2001. Looking Without Seeing: Understanding Unsophisticated Consumers' Success and Failure to Detect Internet Deception. In *ICIS*. 13.

[69] K. Greene, M. Steves, M. Theofanos, and J Kostick. 2018. User Context: An Explanatory Variable in Phishing Susceptibility. (2018).

[70] T. Greening. 1996. Ask and ye shall receive: a study in social engineering. *ACM SIGSAC Review* 14, 2 (1996), 8–14.

[71] R. Griggs and J. Cox. 1982. The elusive thematic-materials effect in Wason's selection task. *Br. J. Psychol.* 73, 3 (1982), 407–420.

[72] S. Gupta, P. Gupta, M. Ahamad, and P. Kumaraguru. 2015. Abusing Phone Numbers and Cross-Application Features for Crafting Targeted Attacks. *arXiv:1512.07330 [cs]* (2015).

[73] S. Gupta and P. Kumaraguru. 2014. Emerging phishing trends and effectiveness of the anti-phishing landing page. In *APWG eCrime*. 36–47.

[74] M. Hale, R. Gamble, and P. Gamble. 2015. CyberPhishing: A Game-Based Platform for Phishing Awareness Testing. In *Hawaii International Conference on System Sciences*. 5260–5269.

[75] X. Han, N. Kheir, and D. Balzarotti. 2016. PhishEye: Live Monitoring of Sandboxed Phishing Kits. In *CCS*. ACM, 1402–1413.

[76] B. Harrison, E. Svetieva, and A. Vishwanath. 2016. Individual processing of phishing emails: How attention and elaboration protect against phishing. *Online Inf. Rev.* 40, 2 (2016), 265–281.

[77] B. Harrison, A. Vishwanath, Y. Ng, and R. Rao. 2015. Examining the Impact of Presence on Individual Phishing Victimization. In *Hawaii International Conference on System Sciences*. 3483–3489.

[78] R. Hastie and R. Dawes. 2010. *Rational Choice in an Uncertain World: The Psychology of Judgment and Decision Making*. SAGE.

[79] R. Heartfield and G. Loukas. 2015. A Taxonomy of Attacks and a Survey of Defence Mechanisms for Semantic Social Engineering Attacks. *ACM Comput. Surv.* 48, 3 (2015), 37:1–37:39.

[80] G. Ho, A. Cidon, L. Gavish, M. Schweighauser, V. Paxson, S. Savage, G. Voelker, and D. Wagner. 2019. Detecting and Characterizing Lateral Phishing at Scale. In *USENIX Security Symposium*. 1273–1290.

[81] H. Holm, W.R. Flores, M. Nohlberg, and M. Ekstedt. 2014. An empirical investigation of the effect of target-related information in phishing attacks. In *International Enterprise Distributed Object Computing Workshop*. IEEE, 357–363.

[82] H. Holm, W. R. Flores, and G. Ericsson. 2013. Cyber security for a Smart Grid - What about phishing?. In *IEEE PES ISGT Europe*. 1–5.

[83] O. Houdé, L. Zago, E. Mellet, S. Moutier, A. Pineau, B. Mazoyer, and N. Tzourio-Mazoyer. 2000. Shifting from the Perceptual Brain to the Logical Brain: The Neural Impact of Cognitive Inhibition Training. *J. Cogn. Neurosci.* 12, 5 (2000), 721–728.

[84] D. House and M.K. Raja. 2019. Phishing: message appraisal and the exploration of fear and self-confidence. *Behav. Inf. Technol.* (2019).

[85] D. Irani, M. Balduzzi, D. Balzarotti, E. Kirda, and C. Pu. 2011. Reverse Social Engineering Attacks in Online Social Networks. In *Detection of Intrusions and Malware, and Vulnerability Assessment (LNCS)*. Springer, 55–74.

[86] C. Iuga, J.. Nurse, and A. Erola. 2016. Baiting the hook: factors impacting susceptibility to phishing attacks. *Hum.-centr. Comput. Inf. Sci* 6, 1 (2016).

[87] C. Jackson, D. Simon, Desney S. Tan, and A. Barth. 2007. An Evaluation of Extended Validation and Picture-in-Picture Phishing Attacks. In *Financial Cryptography and Data Security (LNCS)*. Springer, 281–293.

[88] T. Jagatic, N.. Johnson, M. Jakobsson, and F. Menczer. 2007. Social phishing. *Commun. ACM* 50, 10 (2007), 94–100.

[89] M. Jakobsson and J. Ratkiewicz. 2006. Designing ethical phishing experiments: a study of (ROT13) rOnl query features. In *WWW*. ACM, 513–522.

[90] M. Jakobsson, A. Tsow, A. Shah, E. Blevis, and Y.-K. Lim. 2007. What instills trust? A qualitative study of phishing. In *Usable Security (LNCS, Vol. 4886)*. Springer, 356–361.

[91] M.. Jalali, M. Bruckes, D. Westmattelmann, and G. Schewe. 2020. Why Employees (Still) Click on Phishing Links: Investigation in Hospitals. *J. Med. Internet Res.* 22, 1 (2020).

[92] K. Jansson and R. Solms. 2013. Phishing for phishing awareness. *Behav. Inf. Technol.* 32, 6 (2013), 584–593.

[93] M. Jensen, M. Dinger, R. Wright, and J. Thatcher. 2017. Training to Mitigate Phishing Attacks Using Mindfulness Techniques. *J. Manag. Inf. Syst.* 34, 2 (2017), 597–626.

[94] O. P John, E. Donahue, and R. Kentle. 1991. Big five inventory. *J Pers Soc Psychol* (1991).

[95] M. Junger, L. Montoya, and F.-J. Overink. 2017. Priming and warnings are not effective to prevent social engineering attacks. *Comput. Hum. Behav.* 66 (2017), 75–87.

[96] W.D. Kearney and H.A. Kruger. 2016. Can perceptual differences account for enigmatic information security behaviour in an organisation? *Computers and Security* 61 (2016), 46–58.

[97] B. Kim, D.-Y. Lee, and B. Kim. 2019. Deterrent effects of punishment and training on insider security threats: a field experiment on phishing attacks. *Behav. Inf. Technol.* (2019).

[98] T. Kindberg, E. O'Neill, C. Bevan, V. Kostakos, D.S. Fraser, and T. Jay. 2008. Measuring Trust in Wi-Fi hotspots. In *Conference on Human Factors in Computing Systems*. ACM, 173–182.

[99] S. Kleitman, M. Law, and J. Kay. 2018. It's the deceiver and the receiver: Individual differences in phishing susceptibility and false positives with item profiling. *PLOS ONE* 13, 10 (2018).

[100] KnowBe4. 2021. *Phishing*. Retrieved March 22, 2022 from https://www.knowbe4.com/phishing

[101] Vadim Kotov and Fabio Massacci. 2013. Anatomy of Exploit Kits. In *Engineering Secure Software and Systems (LNC&S)*. Springer, 181–196.

[102] P. Kumaraguru, J. Cranshaw, A. Acquisti, L. Cranor, J. Hong, M.A. Blair, and T. Pham. 2009. School of phish: A real-world evaluation of anti-phishing training. In *Symposium On Usable Privacy and Security*.

[103] P. Kumaraguru, Y. Rhee, S. Sheng, S. Hasan, A. Acquisti, L.F. Cranor, and J. Hong. 2007. Getting users to pay attention to anti-phishing education: Evaluation of retention and transfer. In *International Conference Proceeding Series*, Vol. 269. ACM, 70–81.

[104] P. Kumaraguru, S. Sheng, A. Acquisti, L. Cranor, and J. Hong. 2010. Teaching Johnny not to fall for phish. *TOIT* 10, 2 (2010), 7:1–7:31.

[105] D. Lain, K. Kostiainen, and S. Capkun. 2021. Phishing in Organizations: Findings from a Large-Scale and Long-Term Study. *arXiv:2112.07498* (2021).

[106] R. Langner. 2011. Stuxnet: Dissecting a Cyberwarfare Weapon. *IEEE Secur Priv* 9, 3 (2011).

[107] R. Langner. 2013. *To Kill a Centrifuge - A Technical Analysis of What Stuxnet's Creators Tried to Achieve*. Technical Report.

[108] E. Lastdrager, I.C. Gallardo, P. Hartel, and M. Junger. 2019. How effective is anti-phishing training for children?. In *SOUPS 2019*. USENIX, 229–239.

[109] N. Lavie, A. Hirst, J. de Fockert, and E. Viding. 2004. Load Theory of Selective Attention and Cognitive Control. *J. Exp. Psychol. Gen.* 133, 3 (2004), 339–354.

[110] Y. Li, K. Xiong, and X. Li. 2019. An analysis of user behaviors in phishing email using machine learning techniques. In *International Joint Conference on e-Business and Telecommunications*, Vol. 2. 529–534.

[111] M. Lien, P. Allen, E. Ruthruff, J. Grabbe, R. McCann, and R. Remington. 2006. Visual word recognition without central attention: Evidence for greater automaticity with advancing age. *Psychology and Aging* 21, 3 (2006), 431–447.

[112] E. Lin, S. Greenberg, E. Trotter, D. Ma, and J. Aycock. 2011. Does domain highlighting help people identify phishing sites?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2075–2084.

[113] T. Lin, D.. Capecci, D. Ellis, H. Rocha, S. Dommaraju, D. Oliveira, and N. Ebner. 2019. Susceptibility to Spear-Phishing Emails: Effects of Internet User Demographics and Email Content. *ACM Trans. Comput.-Hum. Interact.* 26, 5 (2019), 32:1–32:28.

[114] X. Luo, W. Zhang, S. Burd, and A. Seazzu. 2013. Investigating phishing victimization with the Heuristic-Systematic model: A theoretical framework and an exploration. *Computers and Security* 38 (2013), 28–38.

[115] Z. Ma, J. Reynolds, J. Dickinson, K. Wang, T. Judd, J.D. Barnes, J. Mason, and M. Bailey. 2019. The impact of secure transport protocols on phishing efficacy. In *USENIX Workshop on Cyber Security Experimentation and Test*.

[116] K. Marett and R. Wright. 2009. The effectiveness of deceptive tactics in phishing. In *Americas Conference on Information Systems*, Vol. 4. 2583–2591.

[117] E. McBee, T. Ratcliffe, K. Picho, A. Artino, L. Schuwirth, W. Kelly, J. Masel, C. van der Vleuten, and S. Durning. 2015. Consequences of contextual factors on clinical reasoning in resident physicians. *Adv Health Sci Educ* 20, 5 (2015), 1225–1236.

[118] S. Michie, R. West, R. Campbell, J. Brown, and H. Gainforth. 2014. *ABC of Behaviour Change Theories*. Silverback Publishing.

[119] K. Mitnick and W. Simon. 2003. *The Art of Deception: Controlling the Human Element of Security*. John Wiley & Sons.

[120] D. Miyamoto, T. Iimura, G. Blanc, H. Tazaki, and Y. Kadobayashi. 2016. EyeBit: Eye-Tracking Approach for Enforcing Phishing Prevention Habits. In *International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*. IEEE, 56–65.

[121] K. Molinaro and M.. Bolton. 2018. Evaluating the applicability of the double system lens model to the analysis of phishing email judgments. *Computers & Security* 77 (2018), 128–137.

[122] R. Montañez, E. Golob, and S. Xu. 2020. Human Cognition Through the Lens of Social Engineering Cyberattacks. *Frontiers in Psychology* 11 (2020).

[123] G.D. Moody, D.F. Galletta, and B.K. Dunn. 2017. Which phish get caught An exploratory study of individuals' susceptibility to phishing. *European Journal of Information Systems* 26, 6 (2017), 564–584.

[124] T. Moore and R. Anderson. 2008. How brain type influences online safety. In *Workshop on Security and Human Behaviour*. 8.

[125] P.L. Morgan, E.J. Williams, N.A. Zook, and G. Christopher. 2019. Exploring Older Adult Susceptibility to Fraudulent Computer Pop-Up Interruptions. In *Advances in Intelligent Systems and Computing*, Vol. 782. Springer, 56–68.

[126] P.M.W. Musuva, K.W. Getao, and C.K. Chepken. 2019. A new approach to modelling the effects of cognitive processing and threat detection on phishing susceptibility. *Comput. Hum. Behav.* 94 (2019), 154–175.

[127] S. Muthal, S. Li, Y. Huang, X. Li, A. Dahbura, N. Bos, and K. Molinaro. 2017. A phishing study of user behavior with incentive and informed intervention. In *Annual Cyber Security Summit*.

[128] J.D. Ndibwile, E.T. Luhanga, D. Fall, D. Miyamoto, G. Blanc, and Y. Kadobayashi. 2019. An Empirical Approach to Phishing Countermeasures through Smart Glasses and Validation Agents. *IEEE Access* 7 (2019), 130758–130771.

[129] A. Neupane, N. Saxena, and L. Hirshfield. 2017. Neural Underpinnings of Website Legitimacy and Familiarity Detection: An fNIRS Study. In *International Conference on World Wide Web*. ACM, 1571–1580.

[130] A. Neupane, N. Saxena, J. Maximo, and R. Kana. 2016. Neural Markers of Cybersecurity: An fMRI Study of Phishing and Malware Warnings. *IEEE Trans. Inf. Forensics Secur.* 11, 9 (2016), 1970–1983.

[131] J. Nicholson, L. Coventry, and P. Briggs. 2017. Can We Fight Social Engineering Attacks By Social Means? Assessing Social Salience as a Means to Improve Phish Detection. In *Symposium on Usable Privacy and Security*. 15.

[132] J.s Nicholson, Y. Javed, M. Dixon, L. Coventry, O. Ajayi, and P. Anderson. 2020. Investigating Teenagers' Ability to Detect Phishing Messages. In *European Symposium on Security and Privacy Workshops*. IEEE, 140–149.

[133] B.A. Nosek, C.B. Hawkins, and R.S. Frazier. 2011. Implicit social cognition: From measures to mechanisms. *Trends Cogn. Sci.* 15, 4 (2011), 152–159.

[134] D. Oliveira, H. Rocha, H. Yang, D. Ellis, S. Dommaraju, M. Muradoglu, D. Weir, A. Soliman, T. Lin, and N. Ebner. 2017. 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*. ACM, 6412–6424.

[135] K. Onarlioglu, U. Yilmaz, E. Kirda, and D. Balzarotti. 2012. Insights into User Behavior in Dealing with Internet Attacks.. In *NDSS (2012)*.

[136] K. Parsons, M. Butavicius, P. Delfabbro, and M. Lillie. 2019. Predicting susceptibility to social influence in phishing emails. *Int. J. Hum. Comput. Stud.* 128 (2019), 17–26.

[137] K. Parsons, A. McCormac, M. Pattinson, M. Butavicius, and C. Jerram. 2013. Phishing for the truth: A scenario-based experiment of users' behavioural response to emails. In *IFIP International Information Security Conference*, Vol. 405. 366–378.

[138] K. Parsons, A. Mccormac, M. Pattinson, M. Butavicius, and C. Jerram. 2014. 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)*.

[139] K. Parsons, A. McCormac, M. Pattinson, M. Butavicius, and C. Jerram. 2015. The design of phishing studies: Challenges for researchers. *Computers and Security* 52 (2015), 194–206.
[140] C. Patsakis and A. Chrysanthou. 2020. Analysing the fall 2020 Emotet campaign. *arXiv:2011.06479* (2020).
[141] M Pattinson, C Jerram, K Parsons, A McCormac, and M Butavicius. 2011. Managing Phishing Emails: A Scenario-Based Experiment. In *Human Aspects of Information Security and Assurance*. 74–85.
[142] M. Pattinson, C. Jerram, K. Parsons, A. McCormac, and M. Butavicius. 2012. Why do some people manage phishing e-mails better than others? *Inf. Manag. Comput. Secur.* 20, 1 (2012), 18–28.
[143] T. Pfeiffer, M. Kauer, and J. Röth. 2014. *"A bank would never write that!' A qualitative study on E-mail trust decisions.* Gesellschaft für Informatik e.V.
[144] S. Pfleeger and D. Caputo. 2012. Leveraging behavioral science to mitigate cyber security risk. *Computers & Security* 31, 4 (2012), 597–611.
[145] Phishlabs. 2019. *Phishing Trends and Intelligence Rreport.* Technical Report.
[146] S. Pirocca, L. Allodi, and N. Zannone. 2020. A Toolkit for Security Awareness Training Against Targeted Phishing. In *Information Systems Security*. Springer, 137–159.
[147] A. Pisarchik, V. Maksimenko, and A. Hramov. 2019. From Novel Technology to Novel Applications: Comment on "An Integrated Brain-Machine Interface Platform With Thousands of Channels" by Elon Musk and Neuralink. *J. Med. Internet Res.* 21, 10 (2019).
[148] S. Purkait. 2012. Phishing counter measures and their effectiveness - Literature review. *Inf. Manag. Comput. Secur.* 20, 5 (2012), 382–420.
[149] S. Purkait, S. Kumar De, and D. Suar. 2014. 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.
[150] Rapid7. 2021. *Phishing Awareness Training.* Retrieved March 22, 2022 from https://www.rapid7.com/solutions/phishing-awareness-training/
[151] S. Rezaei. 2021. Beyond explicit measures in marketing research: Methods, theoretical models, and applications. *J. Retail. Consum. Serv.* 61 (2021).
[152] J. Rudoy and K. Paller. 2009. Who can you trust? Behavioral and neural differences between perceptual and memory-based influences. *Frontiers in Human Neuroscience* 3 (2009).
[153] F. Salahdine and N. Kaabouch. 2019. Social Engineering Attacks: A Survey. *Future Internet* 11, 4 (2019), 89.
[154] N. Salkind. 2010. *Encyclopedia of research design.* SAGE. 467–468 pages.
[155] M. Sarter, B. Givens, and J. Bruno. 2001. The cognitive neuroscience of sustained attention: where top-down meets bottom-up. *Brain Res. Rev.* 35, 2 (2001), 146–160.
[156] J. Shaw. 2020. 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.
[157] S. Sheng, M. Holbrook, P. Kumaraguru, L. Cranor, and J. Downs. 2010. Who falls for phish? a demographic analysis of phishing susceptibility and effectiveness of interventions. In *SIGCHI Conference on Human Factors in Computing Systems*. ACM, 373–382.
[158] S. Sheng, B. Magnien, P. Kumaraguru, A. Acquisti, L. Cranor, J. Hong, and E. Nunge. 2007. Anti-Phishing Phil: the design and evaluation of a game that teaches people not to fall for phish. In *Symposium on Usable Privacy and Security*. ACM, 88–99.
[159] M. Shonman, X. Li, H. Zhang, and A. Dahbura. 2018. Simulating phishing email processing with instance-based learning and cognitive chunk activation. In *Brain Informatics (LNAI)*. Springer, 468–478.
[160] M. Silic and A. Back. 2016. The dark side of social networking sites:Understanding phishing risks. *Comput. Hum. Behav.* 60 (2016), 35–43.
[161] M. Silic and P.B. Lowry. 2020. Using Design-Science Based Gamification to Improve Organizational Security Training and Compliance. *J Manag Inf Syst.* 37, 1 (2020), 129–161.
[162] T. Sommestad and H. Karlzen. 2019. A meta-analysis of field experiments on phishing susceptibility. In *Symposium on Electronic Crime Research*.
[163] N. Stembert, A. Padmos, M.S. Bargh, S. Choenni, and F. Jansen. 2016. A Study of Preventing Email (Spear) Phishing by Enabling Human Intelligence. In *European Intelligence and Security Informatics Conference*. IEEE, 113–120.
[164] G. Stivala and G. Pellegrino. 2020. Deceptive Previews: A Study of the Link Preview Trustworthiness in Social Platforms. In *NDSS Symposium*.
[165] S. Stockhardt, B. Reinheimer, M. Volkamer, P. Mayer, A. Kunz, P. Rack, and D. Lehmann. 2016. Teaching Phishing-Security: Which Way is Best?. In *ICT Systems Security and Privacy Protection*. Springer, 135–149.
[166] B. Stout. 2021. *Email Credential Harvesting at Scale Without Malware.* Retrieved March 22, 2022 from https://edu.nl/ur4pv
[167] P. Tetri and J. Vuorinen. 2013. Dissecting social engineering. *Behav. Inf. Technol.* 32, 10 (2013), 1014–1023.
[168] R. Thaler and C. Sunstein. 2009. *Nudge: improving decisions about health.* Penguin.
[169] K. Thomas and S. Clifford. 2017. Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Comput. Hum. Behav.* 77 (2017), 184–197.
[170] Y. Tian, J. Yuan, and S. Yu. 2017. SBPA: Social behavior based cross Social Network phishing attacks. In *CNS*. IEEE, 366–367.
[171] M. Tischer, Z. Durumeric, S. Foster, S. Duan, A. Mori, E. Bursztein, and M. Bailey. 2016. Users Really Do Plug in USB Drives They Find. In *Symposium on Security & Privacy*. IEEE, 306–319.
[172] H. Tu, A. Doupé, Z. Zhao, and G. Ahn. 2019. Users Really Do Answer Telephone Scams. In *USENIX Security Symposium*. 1327–1340.
[173] E. Uhlmann, K. Leavitt, J. Menges, J. Koopman, M. Howe, and R. Johnson. 2012. Getting Explicit About the Implicit: A Taxonomy of Implicit Measures and Guide for Their Use in Organizational Research. *Organ. Res. Methods* 15, 4 (2012), 553–601.
[174] R. Valecha, A. Gonzalez, J. Mock, E. Golob, and H. Raghav Rao. 2020. Investigating Phishing Susceptibility—An Analysis of Neural Measures. In *Information Systems and Neuroscience (LNISO)*. Springer, 111–119.
[175] A. Van Der Heijden and L. Allodi. 2019. Cognitive triaging of phishing attacks. In *USENIX Security Symposium*. 1309–1326.

[176] B. van Dooremaal, P. Burda, L. Allodi, and N. Zannone. 2021. Combining Text and Visual Features to Improve the Identification of Cloned Webpages for Early Phishing Detection. In *International Conference on Availability, Reliability and Security*. ACM, 1–10.
[177] Verizon. 2020. *Data Breach Investigations Report.* Technical Report.
[178] Verizon. 2021. *Data Breach Investigations Report.* Technical Report.
[179] T. Vidas, E. Owusu, S. Wang, C. Zeng, L. Cranor, and N. Christin. 2013. QRishing: The Susceptibility of Smartphone Users to QR Code Phishing Attacks. In *Financial Cryptography and Data Security (LNCS)*. Springer, 52–69.
[180] A. Vishwanath. 2015. Diffusion of deception in social media: Social contagion effects and its antecedents. *Inf. Syst. Front.* 17, 6 (2015), 1353–1367.
[181] A. Vishwanath. 2015. Examining the Distinct Antecedents of E-Mail Habits and its Influence on the Outcomes of a Phishing Attack. *J. Comput.-Mediat. Commun.* 20, 5 (2015), 570–584.
[182] A. Vishwanath. 2015. Habitual Facebook Use and its Impact on Getting Deceived on Social Media. *J. Comput.-Mediat. Commun.* 20, 1 (2015), 83–98.
[183] A. Vishwanath. 2016. Mobile device affordance: Explicating how smartphones influence the outcome of phishing attacks. *Comput. Hum. Behav.* 63 (2016), 198–207.
[184] A. Vishwanath. 2017. Getting phished on social media. *Decis. Support Syst.* 103 (2017), 70–81.
[185] A. Vishwanath, B. Harrison, and Y. Ng. 2016. Suspicion, Cognition, and Automaticity Model of Phishing Susceptibility. *Communication Research* 45, 8 (2016), 1146–1166.
[186] A. Vishwanath, T. Herath, R. Chen, J. Wang, and H. R. Rao. 2011. Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. *Decis. Support Syst.* 51, 3 (June 2011), 576–586.
[187] M. Volkamer, K. Renaud, and B. Reinheimer. 2016. TORPEDO: TOoltip-poweRed Phishing Email DetectiOn. In *ICT Systems Security and Privacy Protection*. Springer, 161–175.
[188] J. Wang, T. Herath, R. Chen, A. Vishwanath, and H. Rao. 2012. Research Article Phishing Susceptibility: An Investigation Into the Processing of a Targeted Spear Phishing Email. *IEEE Trans. Prof. Commun.* 55, 4 (2012), 345–362.
[189] Z. Wang, L. Sun, and H. Zhu. 2020. Defining Social Engineering in Cybersecurity. *IEEE Access* 8 (2020), 85094–85115.
[190] R. Wash and M.M. Cooper. 2018. Who provides phishing training? Facts, stories, and people like me. In *CHI*. ACM, 1–12.
[191] E.. Williams, J. Hinds, and A. Joinson. 2018. Exploring susceptibility to phishing in the workplace. *Int. J. Hum. Comput. Stud.* 120 (2018), 1–13.
[192] E. Williams, P. Morgan, and A. Joinson. 2017. Press accept to update now: Individual differences in susceptibility to malevolent interruptions. *Decis. Support Syst.* 96 (2017), 119–129.
[193] P. Winkielman, K. Berridge, and J. Wilbarger. 2005. Unconscious Affective Reactions to Masked Happy Versus Angry Faces Influence Consumption Behavior and Judgments of Value. *Pers Soc Psychol Bull* 31, 1 (2005), 121–135.
[194] M. Workman. 2007. Gaining access with social engineering: An empirical study of the threat. *Inf. Syst. Secur.* 16, 6 (2007), 315–331.
[195] M. Workman. 2008. A test of interventions for security threats from social engineering. *Inf. Manag. Comput. Secur.* 16, 5 (2008), 463–483.
[196] M. Workman. 2008. 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.
[197] R. Wright, S. Chakraborty, A. Basoglu, and K. Marett. 2010. Where Did They Go Right? Understanding the Deception in Phishing Communications. *Group Decision and Negotiation* 19, 4 (2010), 391–416.
[198] R.T. Wright, M.L. Jensen, J.B. Thatcher, M. Dinger, and K. Marett. 2014. Influence techniques in phishing attacks: An examination of vulnerability and resistance. *Inf. Syst. Res.* 25, 2 (2014), 385–400.
[199] R.T. Wright and K. Marett. 2010. The influence of experiential and dispositional factors in phishing: An empirical investigation of the deceived. *J Manag Inf Syst.* 27, 1 (2010), 273–303.
[200] R.T. Wright, K. Marett, and J.B. Thatcher. 2014. Extending ecommerce deception to phishing. In *International Conference on Information Systems.*
[201] M. Wu, R. Miller, and S. Garfinkel. 2006. Do security toolbars actually prevent phishing attacks?. In *CHI*. ACM, 601–610.
[202] W. Yang, J. Chen, A. Xiong, R.W. Proctor, and N. Li. 2015. Effectiveness of a phishing warning in field settings. In *HotSoS*. ACM, Article 14.
[203] W. Yang, A. Xiong, J. Chen, R. Proctor, and N. Li. 2017. Use of Phishing Training to Improve Security Warning Compliance: Evidence from a Field Experiment. In *Hot Topics in Science of Security: Symposium and Bootcamp*. ACM, 52–61.
[204] F. Zahedi, A. Abbasi, and Y. Chen. 2015. Fake-Website Detection Tools: Identifying Elements that Promote Individuals' Use and Enhance Their Performance. *J. Assoc. Inf. Syst.* 16, 6 (2015).
[205] H. Zhang, S. Singh, X. Li, A. Dahbura, and M. Xie. 2018. Multitasking and Monetary Incentive in a Realistic Phishing Study. In *International Human Computer Interaction Conference*. BCS Learning & Development Ltd.
[206] O. Çetin, C. Gañán, L. Altena, S. Tajalizadehkhoob, and M. van Eeten. 2019. Tell Me You Fixed It: Evaluating Vulnerability Notifications via Quarantine Networks. In *European Symposium on Security and Privacy*. IEEE, 326–339.