Dissecting Social Engineering Attacks Through the Lenses of Cognition

Pavlo Burda
Eindhoven University of Technology
p.burda@tue.nl

Luca Allodi
Eindhoven University of Technology
l.allodi@tue.nl

Nicola Zannone
Eindhoven University of Technology
n.zannone@tue.nl

***Abstract*—In this paper we present, showcase, and analize a novel framework to dissect Social Engineering (SE) attacks. The framework is based on extant theories in the cognitive sciences, and is meant as an instrument for researchers and practitioners alike to structure and analyze SE attacks of varying sophistication, isolating specific features and their effects at the cognitive level, and providing a common structure for comparisons across different attacks. We showcase the framework against attacks reproduced in the academic literature as well as against real (highly-targeted) SE attacks reported in the wild, isolating and relating effects and techniques adopted by the attackers to the target's cognitive process. We discuss implications for research and practice of the proposed framework.**

***Index Terms*—social engineering, cognitive science**

1. Introduction

Social Engineering (SE) attacks are taking an increasingly prominent role in the cyber-security threat landscape, from attacks against individual citizens to attacks threatening societal processes, such the EU election and, recently, the supply chain for the distribution of COVID-19 vaccines.

SE attacks are cognitive attacks aiming at deceiving individuals by exploiting 'vulnerabilities' inborn in human cognition, with the goal of gaining access to confidential information and/or deliver malware on the target's system [1]. In particular, the last years have witnessed an increased sophistication of human-based exploitation techniques, evolving from less sophisticated *'your email account is full, click here to reset your password'* type of attacks to tailored and well targeted attacks exploiting target information [2]. Research in this area must therefore take a multidisciplinary approach to grasp the nuances of the interactions between the technical aspects of an attack, and the cognitive aspects characterizing its human element.

As a result, a new strain of empirical research testing attack features and the corresponding cognitive effects recently emerged. However, empirical research results are often contrasting and hard to contextualize, making it hard to derive effective defensive measures for real world applications. Sommestad and Karlzen [3] provide a recent overview of the direction and size of cognitive effects in SE, but the question of how to develop SE research to support a coherent interpretation of cognitive effects, and test related countermeasures, remains an open problem.

We argue that the primary reason for this is the lack of a structured, shared understanding of the dimensionality of the SE problem within the research community: the multidisciplinarity of the domain makes it particularly difficult to identify gaps, open research questions, and to interpret experimental results [4]. For example, studies often focus on single domains (e.g., technological, human-related or design), but experimental designs capable of isolating interaction effects across domains are hard to devise without a clear framework of the underlying cognitive processes. Similarly, most empirical research results are limited to 'untargeted' attack scenarios, whereas little understanding remains of the nuances of *targeted* attacks and exploitation of target-relevant information (e.g., memories). On the other hand, targeted attacks are becoming increasingly relevant to the overall threat landscape [2], [4], stressing the importance of filling the gap between SE research results and real-world situations.

In this paper we underline that understanding the cognitive processes involved in a SE attack is fundamental to (a) advance the field of empirical and theoretical research in SE by identifying gaps and effect interactions; (b) provide a framework to evaluate and contextualize research results; (c) characterize the SE attack surface to, for example, be able to measure threat levels, or devise research toward effective policies to thwart targeted SE attacks. To this end, we develop and showcase a cognitive framework for characterizing SE attacks based on theories and models of human cognition drawn from the field of cognitive sciences. The framework can support the design of experiments in the SE domain (e.g., by providing a structure to isolate cognitive effects), as well as being employed to characterize and study existing, sophisticated attacks in the wild thus helping uncovering novel attack techniques whose effects may be tested in experimental settings. To showcase the framework's application, we analyze two academic experiments simulating SE attacks [5], [6] and two real SE attack cases [2], [4] to illustrate how the framework can be used to identify gaps and ways forward.

The paper is structured as follows. Section 2 presents the derivation and description of the cognitive framework and Section 3 showcases it. Section 4 discusses open gaps and ways forward. Section 5 discusses related work and Section 6 concludes the paper.

2. Cognitive framework of SE attacks

2.1. Framework derivation

Cognitive sciences identify a general set of components that constitute the architecture of the cognitive processes of the human mind whose body of evidence stems from the fields of psychology, linguistics and neuroscience [8], [9].

TABLE 1: Theories and models extracted from the cognitive science literature.

Theory/Model	References	Key aspects
Working Memory model (WMM)	Baddeley [7], Hastie and Dawes [8], Anderson (pp. 129-131) [9]	This model is a multi-module model where input and output modules encode information from sensory systems. The main module is the `Working Memory` (`WE`) where manipulation of information from *perception* modules occur. A major system is the `Long-Term Memory` (`LTM`) that contains all sorts of information including procedures for thinking and deciding. The controller of these modules is the `Central Executive` (`CE`) which functions as the *attention* selector and controller of explicit and implicit cognition.
Dual-process theories (DpT)	Evans [10] Sanfey et al. [11] De Neys and Glumicic [12] Oppenheimer [13], Hastie and Dawes (pp. 21-27) [8]	Dual-processing models theorize two different modes for cognitive processing, one highly cognitive-intensive, and the other engaging only low-cognition. A common conceptualization of these modes are "System-1 and System-2", where two systems compete for over response: one is unconscious, rapid, automatic, and high capacity; the other is conscious, slow and deliberative. Regardless of the conceptualization, these theories suggest a mixed use of fast and approximate computations based on *heuristics* where only the final product (e.g., behavior) is posted in consciousness. The engagement of higher or lower cognition is mediated by exogenous and endogenous contextual variables, for example, the environment, fluency, etc.
Global Workspace theory (GWT)	Baars [14] Dehaene and Naccache [15] Baars and Franklin [16]	This theory focuses on conscious processing whose coordination and control depend on the `CE` and its access to a global workspace where other processes are running automatically and unconsciously. The cognitive cycle starts with a competition for consciousness of signals from *perception* modules and `LTM`, which then can proceed to the `WM` (via *attention*) if goal-relevant or picked up by other modules of the workspace. *Attention* modulates the access to consciousness. `WM` and `LTM` function as a substrate for the cognitive cycle.
Expected Utility theory (EUT)	Hastie and Dawes (pp. 551-552) [8]	This theory describes a rational decision making method under uncertainty where individuals seek the highest combination of subjective value (utility) and the highest (expected) probability of events. This decision-making technique can be enabled only by perfectly rational agents.
Prospect theory (PT)	Hastie and Dawes (pp. 655-658) [8]	This theory describes a decision making method under uncertainty where individuals seek the highest utility and the highest prospect (the potential to happen in a desired way) of events. The difference with the expected utility theory lies in the asymmetry of weighting the probability of events for which, for example, losses can have bigger weights than equal gains. This theory corroborates the dual-processing nature in dual-process theories.
Load theory of attention (LToA)	Lavie [17]	*Perception* processes all stimuli in an automatic mandatory fashion until capacity permits. In case of high perceptual load it is less easy to get distracted by goal-irrelevant stimuli. Whereas in high cognitive load it is easier to get distracted by goal-irrelevant stimuli. Thus, an effortful goal-relevant *attention* is necessary in directing executive functions (`CE`) and distractors can more easily disrupt goal-relevant processing.
Feature-integration theory (FiT)	Anderson (pp. 62-65) [9]	People typically focus their visual attention on a stimulus before they can synthesize its features into a pattern. This happens in the *perception* step where early perceptual processing occurs and patterns are recognized. It follows that anomalies are easier to spot when their features do not mix well in a perceived pattern. Thus, selective *attention* is needed to perform an array search between similar features which is a more difficult task.

To derive the building blocks of our framework, we investigated popular theories of cognition in the cognitive sciences [8], [9], and mapped those to the SE domain. Table 1 presents an overview of the extant theories and models of cognition summarizing the common perspectives of cognitive processes. We examine the cognitive processes corresponding to human information processing that can be affected by an attacker during an SE attack, as reported by various studies in this domain [18]–[22]. The identified 'building blocks' of the human cognitive processes that are relevant to SE attacks are reported in Table 2, pictured in Figure 1, and presented in detail in the next section.

2.2. Framework building blocks

Stimulus: The stimulus is any input (e.g., an event, a sound, a message) that triggers a cognitive process. In the SE context, the stimulus represents the means by which the attack is delivered to its (human) target. A stimulus is characterized by attributes describing its content and form. Examples of attributes can be presence/absence of a spoofed address in an email [23], style of writing [24], or the presence of text aimed to evoke past memories of the target [25]. These attributes contribute in determining which components of the framework may be "activated" as the information is processed.

Parameters: Parameters are used to capture contextual information during the cognitive process. Context is assumed to influence many aspects of the production and understanding of text and speech, and is defined as the set of subjective constructions or "definitions" of the *relevant dimensions* (i.e., parameters) of social or communicative situations [26]. We distinguish between *attack parameters* and *target parameters*. Attack parameters represent the assumptions that the attacker makes on the targets and their context. Target parameters characterize the properties of the target and the context in which the target is when the external stimulus arrives. Thus, target parameters mediate the processing pipeline from stimulus to behavior and define the overall context in which the cognitive processes takes place. As shown later, this distinction allows us to reason on the level of targetization of an attack and its effectiveness as the success of the attack is strongly related to the alignment of attack parameters with target parameters [27], [28].

Perception: Perception decodes the sensory information from an incoming stimulus. Perception is a complex process spanning from audio-visual interpretation to features integration and pattern recognition. In particular, perception functions as a signal receiver that translates

TABLE 2: Building blocks

Component	Theory	Description
Stimuli	All	Any event or object that stimulates the senses.
Perception	GWT LToA FiT	A signal receiver that translates the stimulus into percepts. It is mediated by other cognitive processes, like `LTM` associations, that can be concepts, procedures and categorizations, e.g., facial features.
Attention	WMM DpT GWT LToA FiT	A set of systems that modulate the access to consciousness. It has a limited capacity whose allocation can be *exogenous* (controlled by the stimulus) or *endogenous* (goal oriented by the `Central Executive`).
Elaboration	GWT DpT PT EUT	A block responsible for reasoning, like making a decision. It evaluates the available information from the loaded percepts and memory. It allocates cognitive resources, e.g. `WM` or `Attention`, based on currents needs.
Anomaly	DpT	A condition when *Elaboration* cannot deal with the computation due to, e.g., wrong or lack of contextual cues, and engages in effortful processing, like consciously directing attention and making use of `WM`.
Heuristic	DpT	A condition in which *Elaboration* block has found a satisficing rule and engages in low effort processing by relying on heuristics to evaluate information and make inferences.
Behavior	All	The output of the process. It is the response of the whole system to the stimuli, like complying or not complying with the request in the stimulus.
Parameters	—	Properties characterizing the context in which the cognitive process occurs.
Substrate		
`Long-Term Memory`	WMM GWT FiT	A memory system where knowledge is held indefinitely. The two main types of memories are stored therein: explicit recollections of factual information and implicit procedural memories.
`Working Memory`	WMM GWT LToA	A limited capacity system allowing the temporary storage (Short-Term Memory) and manipulation of information necessary for complex tasks as comprehension, learning and reasoning.
`Central Executive`	WMM GWT	An attentional control system that voluntarily manipulates the `WM` functions.

the stimulus into a '*percept*' and automatically loads, from Long-Term Memory (`LTM`) further associations, experiences and judgments related to the stimulus and its attributes based on the contextual parameters of the subject [8], [9]. Perception fetches this information and makes them readily available for the upcoming computation, similarly to what *caching* does in computing. In the SE context, perception is relevant with respect to SE attacks relying on priming [21]: 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 1) but may have strong effects on the subject's subsequent decisions [29]. Related concepts and information are stored in `LTM`, that could then be recalled in form of precepts when a new stimulus arrives, potentially conditioning the targets' decisions as detailed above.

Attention: Attention readies the central nervous system to process and respond to stimuli [9], [16], [17]. Attentional systems select information to process at serial bottlenecks and progressively filter irrelevant signals and information [9] (pp. 53-54). Within the contemporary theories of attention, two general models exist on different cognition levels [17] [9] (p. 54-74): '*peripheral attention*' relates to the sensory experience related to visual and auditory signals, whereas '*central attention*' relates to the semantics of the stimulus at a higher level of abstraction. Since we are concerned with higher level processing, i.e. when stimuli have already been pre-processed, we here consider '*central attention*', whose purpose is to select 'lines of thought' and to focus on a task while allowing for interruption by secondary tasks [9] (pp. 69-72). Central attention influences which and how stimuli are processed depending on the current set of goals in a given moment. 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 in the study of Morgan et al. [30] where they set participants' (central) attention to be endogenous or exogenous to react to malicious pop-ups.

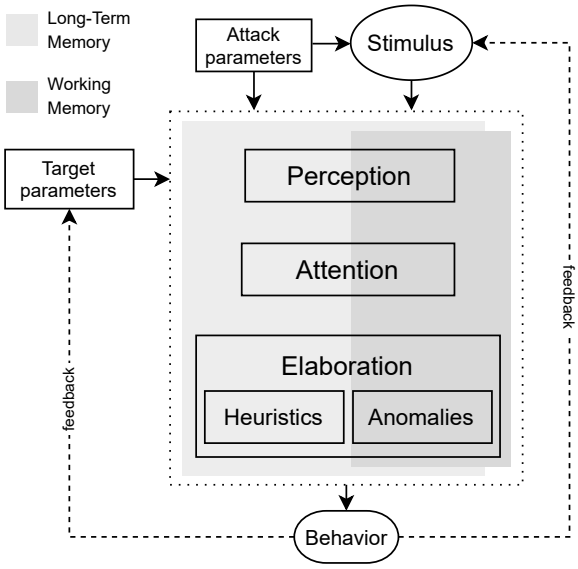

Figure 1: Generic framework of cognition for SE attacks

Elaboration: Elaboration is responsible for processing the information incoming from the other blocks and information stored in memory. The processing involves various conscious and unconscious mental operations performed by a multitude of interconnected and distributed sub-modules [8], [9], [15]. As we are concerned with the cognitive features that can affect cognitive functions with respect to SE attacks, functional and neurobiological definitions of such sub-modules, or the mapping of psychological modules with specific neural circuits, is not of relevance here. Within the scope of this work, the elaboration block is therefore treated as a black box whose operation is influenced by two modulating factors that are known to influence processing and, consequently, behavior in the context of SE: heuristics and anomalies [20], [31], [32].

Heuristics are fast and implicit (that is, not available to introspection) psychological rules that aid judgment and decision making in the elaboration phase [8]. Their use is akin to 'speculative execution' in computing where heuristic processing employs a number of cognitive shortcuts that

lead to appropriate behaviour under most circumstances. Heuristics can emerge from the need of having adequate but fast decisions (e.g., triggering innate behaviour under life-threatening situations), or to lower the cognitive burden associated with repetitive, pattern-specific decisions (e.g., breaking under a red light while driving, or to perform repetitive tasks) [33]–[35]. Cognitive biases such as those described by Cialdini, and often employed in SE research, can also be described as heuristics [20], [36]. Heuristics are stored in `Long-term Memory` and are mostly automatic and unconscious in nature [8]. The effects of heuristics are commonly exploited in all sorts of SE attacks, such as phishing [37] or social networks [38], and are thought to significantly affect the success of attacks [5], [6], [39].

The second influencing factor are *Anomalies*, anomalous conditions that take place when elaboration is unable to handle information that does not fit an automated processing pattern [33], [34] (e.g., a mismatch between URL and the expected domain name). Therefore, the `CE` (see Table 2) has to allocate cognitive resources to accomplish the current task, such as consciously directing attention to the processing of the anomaly, effectively creating a new goal for the elaboration. This requires employing the `WM` to handle the current task and reason on a judgment or decision by means of a wider set cognitive capabilities [15], for example making connections between experiences and knowledge stored in `LTM` with the current case [7]. This mechanism is employed, for example, in anti-phishing training to allow for anomalies to be triggered, where relevant (or "mediating") knowledge is instilled (e.g., what is phishing, what the URL means, etc.) and applied in practice (e.g., embedded phishing exercise) [40]. The availability of relevant knowledge (e.g., expertise) [41], the lack of cognitive resources (e.g., workload, stress) [30] or habits (e.g., context, personality) [31] are all exemplar factors that can condition the triggering of anomalies.

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

Substrate: The substrate represents the computational architecture on top of which the building blocks run [8]. The main components are described at the bottom of Table 2. The cognitive framework operates on a substrate made from the `Long-Term Memory` (`LTM`) and `Working Memory` (shaded in Fig. 1). These two components comprise multiple processing and memory sub-modules and constitute a 'workbench' for mental processes [8], [16]. The `Central Executive` (`CE`, not present in the figure for readability) is responsible for the coordination of mental processes, control of selective endogenous attention and inhibition of automatic responses [16].

3. Cognitive analysis of SE attacks

To illustrate the framework and its applicability to a range of SE scenarios, we apply it to model two SE attacks simulated in academic experiments [5], [6] and two real SE attack cases from the literature [2], [4] of varying 'sophistication'. This illustration showcases how both *real* and *synthetic* SE attacks can be interpreted and broken down using the proposed cognitive framework, similarly to the what is done in [42]. The symbols used through the description of SE attacks are shown in Table 3.

TABLE 3: Notation

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

3.1. Breakdown of two SE academic experiments

Parking fine phishing attack [5]

The first example is derived from a study on phishing susceptibility [5] and represents a simple phishing attempt whose pretext is a parking fine issued by (allegedly) a local police authority. The targets are nudged to click on a link in the phishing email; if this happens, the attack is considered successful. A representation of the attack using our framework is given in Fig. 2 and the verbatim text of the phishing email is provided in Appendix A (Listing 1).

Stimulus & Parameters: The email is the stimulus X triggering the cognitive process of the target. Oliveira et al. [5] explicitly implemented an *authority persuasion technique* in the phishing email (modeled as the attribute of the stimulus γ_1 in Fig. 2) and considered the attack parameters *age* (α_1^p) and *life domain* (α_2^s) in their experiment design. However, other parameters, such income, attention state, car ownership, etc., may also be relevant to the cognitive process of the victim. For simplicity, we include here attack parameters α_1^s=*goal-relevancy:exogenous* (as the stimulus is likely unrelated to the focus of the target when receiving it), and α_2^p=*car owner:true* (as the attacker assumes the target owns a car). These considerations emerge naturally from the attack description and pretext respectively given in [5].

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

Attention: As in the attack simulation run in [5] the subjects do not expect to receive the provided stimulus, the pretext is unlikely to be linked to the current activity of the targets. Therefore, in most instances of the attack the attention block will process the stimulus as 'exogenous' to the current setting, matching the attacker expectation defined in α_1^s. Therefore, the initial elaboration is likely to be influenced by the use of less resource-demanding heuristics.

Elaboration: The attack implements the persuasion technique 'authority' (γ_1), exploiting the associated cognitive bias to increase the chances the target will comply with the email [36]. As shown in [5], authority is particularly effective when related to the legal domain and against young people, which are represented in our framework by attacker parameters α_2^s and α_1^p. Matching these parameters to the actual subjects receiving the stimulus will increase the

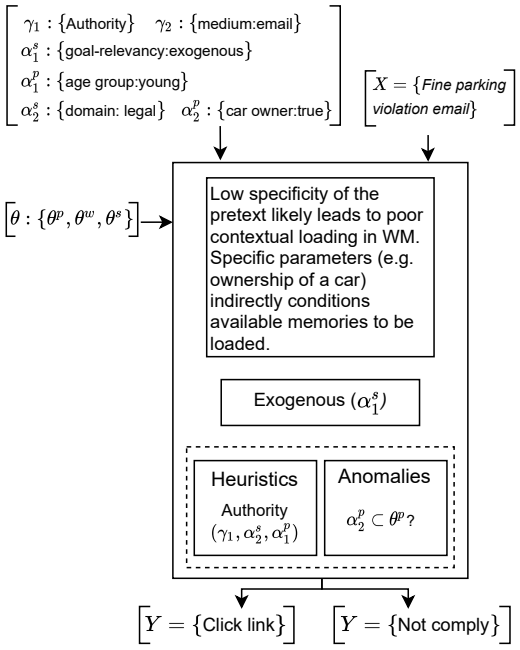

Figure 2: Untargeted fine-parking phishing experiment [5]

chances the target will employ heuristic processing once directed here from the *Attention* block. On the other hand, the elaboration may occur with a higher amount of resources if an anomaly is detected. For example, if the subject does not own a car, i.e., there does not exist a target parameter matching α_2^p, an *Anomaly* is likely to engage more WM during elaboration (akin to re-reading a sentence that does not make sense at a first glance). Additional anomalies may be caused by the detection of a 'suspicious' URL in the email (e.g., as influenced by a subject's technical knowledge, a possible parameter in θ^p), or of an unknown sender.

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

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

Tailored phishing against organizations [6]

The second example concerns a study on tailored phishing against a university and a consultancy company where a bogus organization department asks employees to update their holiday schedule. The pretexts are carefully designed to mimic internal communication patterns and cognitive exploits are employed to enhance the efficacy of the attack [6]. The targets are nudged to click on a link

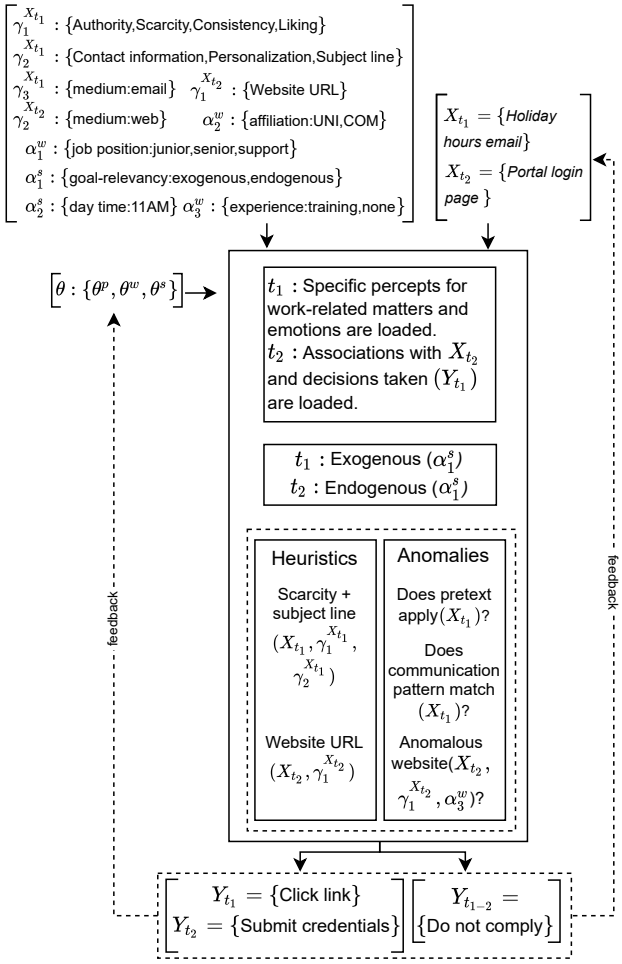

Figure 3: Tailored phishing experiment [6]

and enter their credentials on the fake company page; only submissions to the fake portal are considered as successful, making this a two stage attack. To capture this, we represent the stage in which a stimulus is used and write X_{t_i} to denote the stimulus used in the i-th stage of the attack. A representation of the attack using our framework is given in Fig. 3 and the verbatim text of the phishing email is provided in Appendix B (Listing 2).

Stage 1

Stimulus & Parameters: The email is the first stimulus triggering the cognitive process of the target (X_{t_1} in Fig. 3). [6] tests four persuasion techniques: Authority, Scarcity, Consistency and Liking, represented as an attribute of the stimulus ($\gamma_1^{X_{t_1}}$). Additionally, every persuasion technique is enhanced with three notification methods: extended Contact information, Personalization towards the target and extended Subject line ($\gamma_2^{X_{t_1}}$ in Figure 3).

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

of a work day). These considerations emerge from the experiment design and description given in [6].

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

Attention: No assumptions on the subjects expecting or not expecting to deal with updating their holidays schedule in that time frame are provided in [6]. However, the pretext is unlikely to be linked to the current activity of the targets as scheduling holidays is a sporadic activity. Therefore, in most instances of the attack, the attention block will process the stimulus as 'exogenous' to the current setting. Hence, a low amount of cognitive resources is likely to be allocated to the initial elaboration of the stimulus.

Elaboration: This attack implements various persuasion techniques ($\gamma_1^{X_{t_1}}$), which exploit the associated cognitive biases to push the target to complete the decision-making process heuristically [36]. Following the experiment design in [6], the effect of these heuristics is enhanced by their placement in the email (e.g., subject line, contact information, or signature), $\gamma_2^{X_{t_1}}$. On the other hand, the elaboration may occur with a higher amount of cognitive resources if an anomaly is detected, for example, when the subject has already completed a vacation schedule or the pretext does not apply to the target at all. For instance, as reported in [6], interns were 'immune' to the pretext due to their temporary position. Additional anomalies may be caused by inconsistencies with the usual communication patterns at the organization (e.g., as influenced by the subject's experience, e.g. θ^w: senior).

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

Stage 2

Stimuli & Parameters: The second-stage begins with the website (X_{t_2}) led to by the link in the email. We consider the page URL ($\gamma_1^{X_{t_2}}$) as a relevant attribute for the processing the this new stimulus, since the attackers actively masqueraded the URL to look like legitimate [6]. The security knowledge of an employee can be represented with α_3^w to represent whether training has been administered.

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

Elaboration: Although endogenous control is exerted, the exact replica of the page layout and design should accommodate the heuristic processing as the user may be habituated to login to the organization's portal [43]. However, an anomaly can be generated by processing the URL bar of the browser, or due to discrepancies with any other relevant previous percept or memory regarding the present stimulus (e.g., page formatting, 'lock' in the url bar, etc.). On the other hand, these effects also depend on previous knowledge of the user regarding general Internet security practices, possibly as influenced by received training (α_3^w).

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

Discussion: Unlike the previous case, the examined SE attack implements a tailored context for the targets in terms of a higher amount of conceivably matching work parameters. The investigators employ a set of persuasion techniques and delivery methods to favour heuristics-driven elaboration and increase the odds of success. The framework's explicit representation of anomalies allows to reason on the effects of a highly specific pretext on targets' cognitive processes: the α_2^w parameter (university vs. company) leads to different outcomes with respect to α_1^w (junior vs. senior vs. support) where the lack of knowledge of internal organization processes makes junior employees less capable in identifying anomalies in the communication [6]. Further, the two-stage break down of the simulated attack can make it easier to isolate factors that may not have been considered during the experiment design: what is the influence of a mimicked URL versus a random one with the chosen pretext and parameters?

3.2. Breakdown of two real SE attacks

NGO spear-phishing [2]

This example is a tailored spear phishing attack against an NGO, where the email replays an actual announcement about a conference in Geneva and was edited by the attacker to indicate that all fees would be covered by the organizers and encourages to open an attachment [2]. A representation of the attack using our framework is given in Fig. 4 and the verbatim text of the phishing email is provided in Appendix C (Listing 3).

Stimulus & Parameter: The email pretext clearly revolves around the human rights topic in the context of NGOs and is anchored to two specific themes regarding the Uyghur population and a conference in Geneva. In particular, the attacker assumes the subject's work parameters (α^w) to map the professional context at a given NGO, and the setting parameters to represent the assumed attentional state (α_1^s), the conference date (α_2^s) and the time of the day when the email is sent, to reflect working hours (α_3^s). The stimulus' attributes comprise the impersonation of the sender (γ_1) and an invitation with covered costs (γ_2) (a trigger for the reciprocity persuasion technique commonly used in advertisement [36]).

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

Attention: We have no information on whether the targets are focused on actions related to that specific stimulus when it is processed; from the discussion provided in [2], we assume exogenous attention for most targets, which fosters the use of *Heuristics* later during elaboration in Fig. 4.

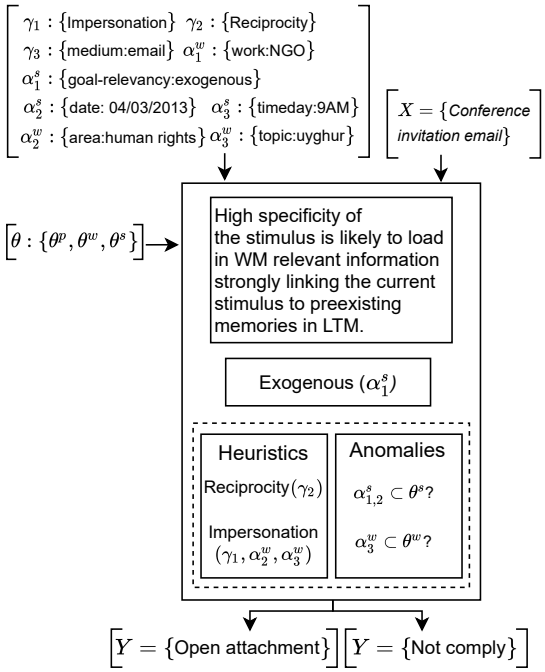

Figure 4: NGO spear-phishing attack [2]

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

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

Discussion: A critical feature of this SE attack is the specificity of the pretext in relation to the experience of the targets. For example, a match of attack parameter α_2^s (i.e., whether the target has registered to the conference) with the actual experience of the target would likely positively reinforce the heuristic judgment. Importantly, were these parameters wrongly calibrated by the attacker, an anomaly would be likely triggered, causing more resources to employed for processing, thus thwarting the attack altogether. The attack flow shows that, unlike the first example (Sec. 3.1), a tailored pretext requires a large set of baseline parameters aligned with the target's context to enable the attack in the first place, similarly to what is discussed in [27]. When this necessary requirement is achieved, the attacker can further develop the attack (e.g., including cognitive exploits like in the second example, Sec. 3.1), to keep the victim's processing anchored to *Heuristics*. Whereas a large attack parameter space increases chances of success when well calibrated, the framework suggests

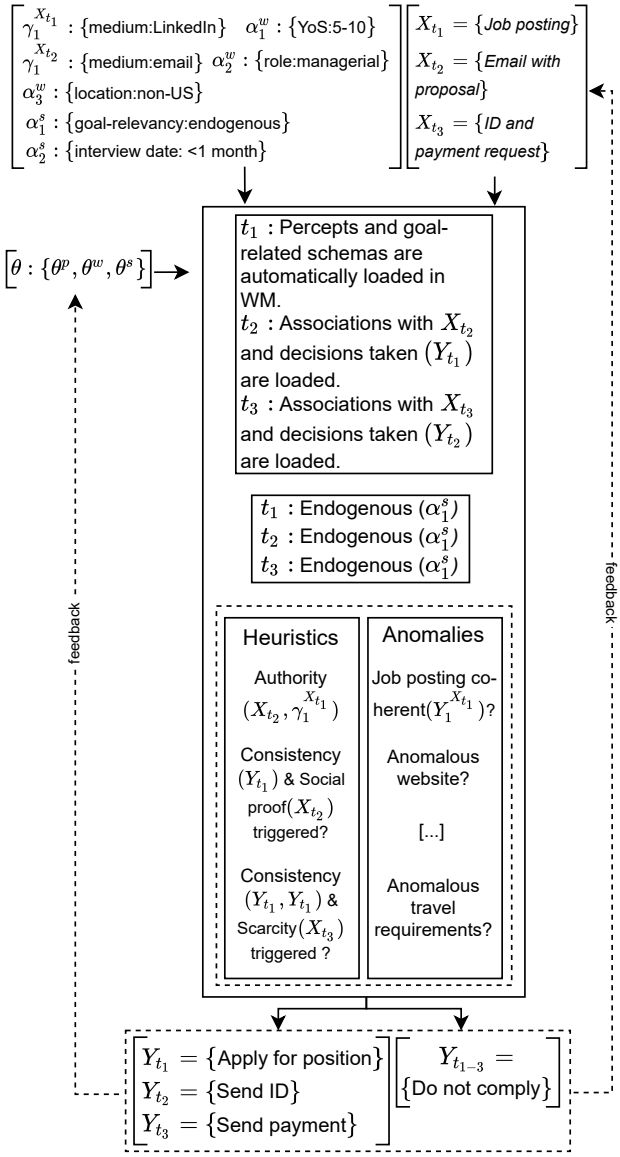

Figure 5: LinkedIn multi-stage attack [4]

that this also increases chances of mismatch, which may backfire and lead to attack failure. With this representation at hand, our framework can potentially enable the design of an ordinal metric to sort similar attacks in terms of matching parameters, amount of knowledge on the targets, akin to [44], and usage of cognitive exploits, e.g. [45].

LinkedIn multi-stage attack [4]

The last example is the case of a highly-targeted spear-phishing campaign against non-US white collar workers on LinkedIn [4], who are offered an appealing job position in the US. Prospect candidates applying for the job are first asked to provide documents and personal details (including a copy of their passport for VISA reasons), and then a payment for the anticipated (fake) traveling costs. Fig. 5 shows the application of the model and the attacker messages are provided in Appendix D (Fig. 6 and Listing 4). This attack evolves through three stages (t_{1-3}), in which different messages are exchanged with the target.

Stage 1

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

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

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

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

Stage 2

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

Perception & Attention: When stimulus X_{t_2} is processed in *Perception* (at t_2 in Fig. 5), previous experience, decisions and associated judgments (Y_{t_1}) are recalled from LTM. These are key aspects to foster deception in this and later stages of the attack as they produce reinforced schemas based on previous experiences that the target will rely on to form upcoming judgments and decisions. The stimulus is still goal-related and, thus, endogenous attention is allocated, leading to a higher usage of WM.

Elaboration: The loading of the percepts in the previous step enables a set of heuristics related to the previous, implicit commitment made in the first attack stage. This is well aligned with the attacker's objective to keep the processing as much as possible towards using *Heuristics*, where X_{t_2} can exploit a number of cognitive biases. The foremost bias exploited in X_{t_2} pushes the subject to remain consistent with their previous decisions (Y_{t_1}) [36]. Additionally, Social proof, Scarcity and Authority can be exploited by the attacker, with the latter two supported by attack parameter α_2^s (i.e., little time ahead of the interview), and stimulus attribute $\gamma_1^{X_{t_1}}$ (i.e., a trustworthy source) respectively. At this point, unless an anomaly is triggered, the heuristic processing reaches a decision whether to continue with the application.

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

Stage 3

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

Discussion: We showed how the model allows one to consistently break down complex attacks into essential steps characterizing the target's cognitive processing of the attack. By highlighting the interaction of multiple stages, we can study the effects of the attacker's strategy, such as the trade-off between target commitment and (escalating) attacker requests (e.g., to define when best to advance a payment request as opposed to asking for additional personal details). We also observe the tactics used to elicit new information, such as applying the Social proof persuasion technique in the second stage to gain a stronger 'foot hold' on the target's side (i.e., it is usual business for large companies to arrange travel and ask for documents). It is worth noting that these considerations are in line with the art of deception whose aim is to reduce suspicion in the target's mind [1], [46]. Importantly, studying the means by which a target's processing flow can be deviated from the processing 'desired' by the attacker may open the way to new training techniques or decision support systems, for example actively detecting the 'escalating' nature of complex SE attacks as revealed by the proposed framework.

4. Discussion

In this paper we propose a theoretical framework of cognition for SE attacks that can guide the formulation, design, and interpretation of SE research. We illustrate the framework application against two real and two simulated SE attacks of different levels of sophistication, and showed how the framework allows one to consistently break down attacks into essential steps characterizing the target's cognitive processing, with different degrees of complexity. This not only allows one to compare different SE attacks based on their cognitive features, but also allows one to reason over why or how was the attack (in-)effective in triggering the target to compliance.

Implications for research. The proposed framework can be used by researchers to systematically identify shortcomings of simulated attacks and experiments, such as isolating factors that are difficult to recognize without a reference to the features of human cognition (e.g., spotting anomalies), and when constructing pretexts and keeping track of targets' context (e.g., the matching of parameters). The framework aids research over several dimensions:

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

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

Attack execution and effect measurement. The framework helps in identifying the key modulating aspects impacting the execution of the attack, for example what type of central attention is expected in the subject when delivering the artifact. The iteration and modifications of attack/subject parameters in multi-stage attacks can also be 'modelled' following the proposed framework: whereas no formal empirical work has been carried out to date on this aspect, the conceptualization proposed by our model shows that, as in the LinkedIn attack example, multiple target-attacker interactions can significantly modify the parameter space across attack stages. The framework can further help researchers in structuring post-experiment measurements, for example by means of surveys, to assess the effects of the attack/stimulus at the different levels of a target's cognitive processes. These include the usage of heuristics, detection of anomalies, but also the possible presence of percepts and memories that affect the computation.

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

Implications for practice. The application of the framework to sophisticated attacks against NGOs and LinkedIn users revealed that the analysis of complex attacks can be simplified and structured to analyze and compare different attacks, their techniques, and execution conditions. Breaking down sophisticated attacks helps to get insights on the causes behind their effectiveness and to devise new detection methods. From a threat intelligence point of view, the forced identification of parameters of an attack, and the match thereof, can help devise risk metrics for different typologies of attacks, for example, based on their level of sophistication. Identifying parameters of an attack may be especially relevant when dealing with internal threats, like ex-employees or undetected compromised accounts, since insider information can be exploited to carry out effective attacks [47]. Importantly, the breakdown of real attacks allows researchers and practitioners alike to keep track of innovative or previously unseen attack techniques, contextualizing and isolating those in the overall cognitive process, opening the way to better training, policies, and research targeted at measuring related effects. Finally, our framework can be useful when designing security systems to reduce the opportunity for cognitive shortcomings to trigger undesired behaviour, similarly to procedures used in design and human computer interaction [42].

5. Related Work

Cognitive processes: Cranor [42] propose a framework for reasoning about the human in the loop to analyze the root cause of security failures attributed to human error. Their framework is based on the Communication-Human Information Processing (C-HIP) model from warning science literature [48]. Pfleeger and Caputo [49] survey behavioral science findings relevant to cyber-security, which partially cover cognitive process features, for example, elaboration and behavior. Montañez et al. [50] map existing studies on various aspects of SE attacks into a basic and selective framework of human cognition functions, and delimit their considerations to stimuli engineering, short and long-term cognitive factors and attention selection aspects till behavior. Steinmetz et al. [46] examine the attributes of SE attacks ascribed by social engineers and reveal that the SE deceptions are intractably intertwined in situational, cultural, and structural circumstances.

Characterization of SE attacks: Heartfield and Loukas [51] propose a taxonomy of semantic SE attacks along with their characteristics and review defense techniques. Tetri and Vourinen [19] introduce a conceptual framework for SE that touches attack characteristics, parameters of targets and setting, and the execution SE attacks. Sommestand and Karlzen [3] analyze phishing field experiments by looking at experimental variables, results and experiment design features, like hypotheses, control variables, etc. The only previous work that proposes a cognitive framework applied to SE attacks is by Cranor [42]. Similarly to our work, the presented framework illustrates the processing of information by a human receiver whose behavior is dependent on a set of processing steps, personal characteristics and environmental disturbances. However, the scope of this work is to facilitate the analysis of secure systems that rely on humans, like an anti-phishing tool providing warnings that may not be heeded by the human. Instead, our framework aim to contextualize SE attacks from the point of view of cognitive sciences where the received inputs and processing steps are related to the attack itself. The work of Steinmetz et al. [46] overlaps with ours in the intention to understand the fundamentals of SE attacks' success, but only from an inherently social psychology perspective.

The interaction of cognitive and social processes described therein naturally lends itself to further deployments of our framework. The other articles that focus on human behavior and cognition aspects [19], [49], [50] do not relate those aspects to SE attacks, or do so implicitly or partially [3], [51]. By contrast, attack characteristics and cognitive processes are the central cornerstone of the present work.

6. Conclusion

In this paper we presented a novel cognitive framework to dissect and characterize social engineering attacks, and relating effects and attack techniques to specific cognitive features and processes of the targets. We showcased the proposed framework against four attacks (from realistic to real, and from general to highly-targeted), illustrating its application both for experimental design/empirical SE research, and as an instrument to characterize attacks in the wild. Future work can focus on the design of metrics to, for example, quantify the sophistication or targetization of an attack employing the proposed framework, or the design of experiments to quantify/verify hypothesized effects on specific cognitive blocks.

Acknowledgment. This work is supported by the ITEA3 programme through the DEFRAUDIfy project funded by Rijksdienst voor Ondernemend Nederland (grant no. ITEA191010). As part of the open-review model followed on WACCO this year, all the reviews for this paper are publicly available at `https://github.com/wacco-workshop/WACCO`.

References

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

[2] S. L. Blond, A. Uritesc, C. Gilbert, Z. L. Chua, P. Saxena, and E. Kirda, "A Look at Targeted Attacks Through the Lense of an {NGO}," 2014, pp. 543–558.

[3] T. Sommestad and H. Karlzen, "A meta-analysis of field experiments on phishing susceptibility," in *APWG eCrime*, 2019.

[4] L. Allodi, T. Chotza, E. Panina, and N. Zannone, "The Need for New Antiphishing Measures Against Spear-Phishing Attacks," *IEEE Security Privacy*, vol. 18, no. 2, pp. 23–34, 2020.

[5] D. Oliveira, H. Rocha, H. Yang, D. Ellis, S. Dommaraju, M. Muradoglu, D. Weir, A. Soliman, T. Lin, and N. Ebner, "Dissecting Spear Phishing Emails for Older vs Young Adults," in *CHI Conference*. ACM, 2017, pp. 6412–6424.

[6] P. Burda, T. Chotza, L. Allodi, and N. Zannone, "Testing the Effectiveness of Tailored Phishing Techniques in Industry and Academia: A Field Experiment," in *ARES*. ACM, 2020.

[7] A. Baddeley, "The episodic buffer: A new component of working memory?" *Trends in Cogn. Sci.*, vol. 4, no. 11, pp. 417–423, 2000.

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

[9] J. R. Anderson, *Cognitive psychology and its implications*. Worth publishers, 2000.

[10] J. Evans, "In two minds: Dual-process accounts of reasoning," *Trends in Cognitive Sciences*, vol. 7, no. 10, pp. 454–459, 2003.

[11] A. Sanfey, G. Loewenstein, S. McClure, and J. Cohen, "Neuroeconomics: Cross-currents in research on decision-making," *Trends in Cognitive Sciences*, vol. 10, no. 3, pp. 108–116, 2006.

[12] W. De Neys and T. Glumicic, "Conflict monitoring in dual process theories of thinking," *Cognition*, vol. 106, pp. 1248–1299, 2008.

[13] D. Oppenheimer, "The secret life of fluency," *Trends in Cognitive Sciences*, vol. 12, no. 6, pp. 237–241, 2008.

[14] B. Baars, "The conscious access hypothesis: Origins and recent evidence," *Trends in Cogn. Sci.*, vol. 6, no. 1, pp. 47–52, 2002.

[15] S. Dehaene and L. Naccache, "Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework," *Cognition*, vol. 79, no. 1-2, pp. 1–37, 2001.

[16] B. Baars and S. Franklin, "How conscious experience and working memory interact," *Trends in Cogn. Sci.*, vol. 7, pp. 166–172, 2003.

[17] N. Lavie, "Distracted and confused?: Selective attention under load," *Trends in Cognitive Sciences*, vol. 9, no. 2, pp. 75–82, 2005.

[18] F. Salahdine and N. Kaabouch, "Social Engineering Attacks: A Survey," *Future Internet*, vol. 11, no. 4, p. 89, 2019.

[19] P. Tetri and J. Vuorinen, "Dissecting social engineering," *Behaviour & Information Technology*, vol. 32, no. 10, pp. 1014–1023, 2013.

[20] A. Ferreira, L. Coventry, and G. Lenzini, "Principles of Persuasion in Social Engineering and Their Use in Phishing," in *Human Aspects of Inf. Sec., Privacy, and Trust*. Springer, 2015, pp. 36–47.

[21] M. Junger, L. Montoya, and F.-J. Overink, "Priming and warnings are not effective to prevent social engineering attacks," *Computers in Human Behavior*, vol. 66, pp. 75–87, 2017.

[22] A. Burns, M. Johnson, and D. Caputo, "Spear phishing in a barrel: Insights from a targeted phishing campaign," *Journal of Org. Computing and Electronic Comm.*, vol. 29, no. 1, pp. 24–39, 2019.

[23] K. Marett and R. Wright, "The effectiveness of deceptive tactics in phishing," in *Americas Conf. on Inf. Sys.*, vol. 4, 2009, pp. 2583–2591.

[24] X. Luo, W. Zhang, S. Burd, and A. Seazzu, "Investigating phishing victimization with the Heuristic-Systematic model: A theoretical framework and an exploration," *Comp. & Sec.*, pp. 28–38, 2013.

[25] R. Wright and K. Marett, "The influence of experiential and dispositional factors in phishing: An empirical investigation of the deceived," *J. of Manag. Inf. Sys.*, vol. 27, pp. 273–303, 2010.

[26] T. A. v. Dijk, "Context and cognition," in *Discourse and Context: A Sociocognitive Approach*. Cambr. Uni. Press, 2008, pp. 56–110.

[27] K. Greene, M. P. Steves, M. F. Theofanos, and J. A. Kostick, "User Context: An Explanatory Variable in Phishing Susceptibility," 2018.

[28] S. Goel, K. Williams, and E. Dincelli, "Got phished? Internet security and human vulnerability," *Journal of the Association for Information Systems*, vol. 18, no. 1, pp. 22–44, 2017.

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

[30] P. Morgan, E. Williams, N. Zook, and G. Christopher, "Exploring Older Adult Susceptibility to Fraudulent Computer Pop-Up Interruptions," *Adv. in Intl. Sys.&Comp.*, vol. 782, pp. 56–68, 2019.

[31] A. Vishwanath, T. Herath, R. Chen, J. Wang, and H. R. Rao, "Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model," *Decision Support Systems*, vol. 51, no. 3, pp. 576–586, 2011.

[32] B. Harrison, E. Svetieva, and A. Vishwanath, "Individual processing of phishing emails: How attention and elaboration protect against phishing," *Online Inf. Review*, vol. 40, no. 2, pp. 265–281, 2016.

[33] D. Kahneman, *Thinking, Fast&Slow*. Farrar, Straus&Giroux, 2011.

[34] ——, "A perspective on judgment and choice: Mapping bounded rationality," *American psychologist*, pp. 697–720, 2003.

[35] A. Tversky and D. Kahneman, "Judgment under Uncertainty: Heuristics & Biases," *Science*, vol. 185, pp. 1124–1131, 1974.

[36] R. Cialdini, *Influence: The Psychology of Persuasion*. Harper Business, 1984.

[37] M. Workman, "Gaining access with social engineering: An empirical study of the threat," *Inf. Sys. Sec.*, vol. 16, pp. 315–331, 2007.

[38] A. Vishwanath, "Getting phished on social media," *Decision Support Systems*, vol. 103, pp. 70–81, 2017.

[39] R. Wright, M. Jensen, J. Thatcher, M. Dinger, and K. Marett, "Influence techniques in phishing attacks: An examination of vulnerability and resistance," *Information Systems Research*, vol. 25, no. 2, pp. 385–400, 2014.

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

[41] J. Wang, T. Herath, R. Chen, A. Vishwanath, and H. R. Rao, "Research Article Phishing Susceptibility: An Investigation Into the Processing of a Targeted Spear Phishing Email," *IEEE Transactions on Professional Communication*, vol. 55, no. 4, pp. 345–362, 2012.

[42] L. F. Cranor, "A Framework for Reasoning About the Human in the Loop," in *USENIX Security Symposium*, 2008.

[43] A. Vishwanath, "Habitual Facebook Use and its Impact on Getting Deceived on Social Media," *Journal of Computer-Mediated Communication*, vol. 20, no. 1, pp. 83–98, Jan. 2015.

[44] S. Pirocca, L. Allodi, and N. Zannone, "A Toolkit for Security Awareness Training Against Targeted Phishing," in *Information Systems Security*. Springer, 2020, pp. 137–159.

[45] A. Van Der Heijden and L. Allodi, "Cognitive triaging of phishing attacks," in *USENIX Security Symposium*, 2019, pp. 1309–1326.

[46] K. F. Steinmetz, A. Pimentel, and W. R. Goe, "Performing social engineering," *Computers in Human Behavior*, vol. 124, 2021.

[47] I. Agrafiotis, J. R. Nurse, O. Buckley, P. Legg, S. Creese, and M. Goldsmith, "Identifying attack patterns for insider threat detection," *Comp. Fraud & Sec.*, vol. 2015, no. 7, pp. 9–17, 2015.

[48] V. C. Conzola and M. S. Wogalter, "A Communication–Human Information Processing (C–HIP) approach to warning effectiveness in the workplace," *J. of Risk Res.*, vol. 4, no. 4, pp. 309–322, 2001.

[49] S. L. Pfleeger and D. D. Caputo, "Leveraging behavioral science to mitigate cyber security risk," *Computers & Security*, vol. 31, no. 4, pp. 597–611, 2012.

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

[51] R. Heartfield and G. Loukas, "A Taxonomy of Attacks and a Survey of Defence Mechanisms for Semantic Social Engineering Attacks," *ACM Computing Surveys*, vol. 48, no. 3, pp. 37:1–37:39, 2015.

Appendix A.
Parking fine phishing attack

This attack is taken from a study on phishing susceptibility [5] and it is a simple phishing attempt which pretext is a parking fine pretending to be from a local police authority. The e-mail content is reported in Listing 1.

Listing 1: Parking fine phishing attack from [5].

```
Our resources have indicated that you have a
parking violation from 12/17/2015 at SW 89th
Avenue at 3:34pm.
Please go to our website to obtain more
information about the violation and to pay your
fine or refute your ticket: <link>
```

Appendix B.
Tailored phishing attack

This attack is taken from an experiment on tailored phishing susceptibility [6] where the authors administer treatments in randomized fashion to employees of a university and a consultancy company. The (first stage) e-mail content is reported in Listing 2. The second stage is a replica of the organization's intranet login page hosted on a mimicked domain name.

Listing 2: Tailored phishing attack against organizations [6]

```
From: info@{domain-name}
Subject: Your holiday hours
Dear Colleague,

To facilitate the planning of activities for the
period September to December, we invite you to
provide a rough estimate of the holiday hours
you are currently planning to take until the end
of this calendar year.
Please provide this information by following
this link:
{domain-name/path}

Thank you,
{signature}
```

Appendix C.
NGO spear-phishing attack

This attack is an advanced spear phishing attack against an NGO [2]. The topic and wording is targeted to the victims, the pretext refers to real specific events that are of interest to the victims and impersonation of high-profile identities is attempted too (with different techniques like spoofing or typos, omitted in the verbatim text). The e-mail content is reported in Listing 3.

Listing 3: NGO spear phishing attack from [2].

```
From: ...
Date: Mon, Mar 4, 2013 at 8:58 AM
Subject: Invitation Letter of WUC International
Conference
To: ...
Dear ...,

I am writing to you from the World Uyghur
Congress (WUC) and on behalf of the
Unrepresented Nations and Peoples Organization (
UNPO) and the Society for Threatened People (STP
) with financial support from the National
Endowment of Democracy, cordially invites you to
 attend the WUC's upcoming Conference which will
 be held in Geneva between 11th and 13th March
2013.

Attached you can find the invitation letter. We
hope you will give a positive consideration to
this invitation, and look forward to meeting you
 in Geneva. During your stay in Geneva, travel,
accommodation and food are covered by the WUC.

The WUC is a non-profit organization granted by
the National Endowment for Democracy in
Washington, DC to peacufully promote human
rights, democracy and freedom for the Uyghur
people in East Turkestan.
If you have any questions or queries regarding
your participation, please do not hesitate to
contact me. Phone: ..., Fax: ..., e-mail: ...

sincerely,
```

Appendix D.
LinkedIn multi-stage attack

This attack is a multi-stage, highly targeted spear-phishing attack against white-collar workers on LinkedIn [4] that actively employs collected information on its

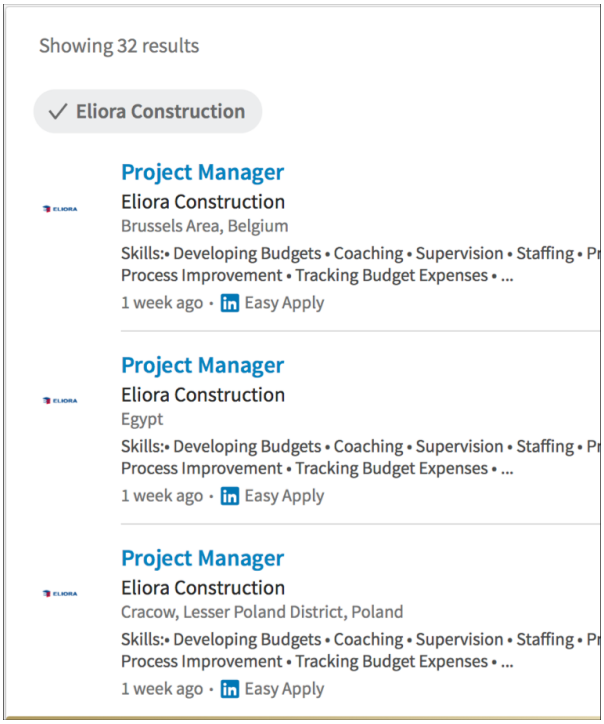

Figure 6: Stage 1 - The LinkedIn job post.

targets to forge the attack artifacts used in each stage of the attack. The LinkedIn post in Fig. 6 refers to a (fictitious) Eliora Construction company located in the US. The offer is targeted towards a specific set of European, North African and Middle East countries where white-collar workers may be more easily appealed to it. Listing 4 presents relevant portions of the artifacts used in the next stages of the attack.

Listing 4: LinkedIn multi-stage attack from [4].

```
[STAGE 2]

Dear Applicant,
I write to inform you that your resume has been
properly reviewed and screened by our recruiting
 board and you have been found eligible for this
 vacant position. Be informed that you have been
 shortlisted for an interview scheduled for
Friday, 12th of January 2018 at ELIORA
CONSTRUCTION COMPANY, 1055 Metropolitan Avenue
Charlotte, North Carolina, 28204, United States
of America.

[...] our primary reason for requesting for your
 physical presence is to have our chief project
manager have a one on one interview with you and
 ensure you possess the aforementioned qualities
 and also have you familiarize yourself with the
 company structure as well as a recap on past
and upcoming project.

[...] Please note that our official travelling
consultant shall handle your travel needs which
will include flight tickets, hotel reservations,
 visa procurement and transfers within the
United States. More so, you will be responsible
for all your travel expenses made through our
affiliated travel agency. These expenses shall
then be refunded to you by Eliora Construction
on arrival at the interview venue

[STAGE 3]

[...] Job Locations: As advertised on LinkedIn (
Further information will be issued after the
interview).

[...] Our company's accountant will furnish you
with our banking details for making a wire
transfer of your booking cost as soon as your
documents have been received.

[...] Interviews are also designed to ascertain
claims of working experience. Should any claim
be found wanting the affected expatriate may be
deported. Please note that our official
travelling consultant shall handle your travel
needs

[...] you will be responsible for all your
travel expenses made through our affiliated
travel agency.

[...] Date of Interview: Friday, 12th of January
 2018
```