Economics Working Paper Series

2018/024

Many Phish in the C: A Coexisting-Choice-Criteria Model of Security Behavior

Iain Embrey and Kim Kaivanto

The Department of Economics
Lancaster University Management School
Lancaster LA1 4YX
UK

LUMS home page: http://www.lancaster.ac.uk/lums/

Many Phish in the $\mathcal{C}$:
A Coexisting-Choice-Criteria Model of Security Behavior*

Iain Embrey† and Kim Kaivanto
Department of Economics, Lancaster University, Lancaster LA1 4YX, UK

this version: November 15, 2018

Abstract

Normative decision theory proves inadequate for modeling human responses to the social-engineering campaigns of Advanced Persistent Threat (APT) attacks. Behavioral decision theory fares better, but still falls short of capturing social-engineering attack vectors, which operate through emotions and peripheral-route persuasion. We introduce a generalized decision theory, under which any decision will be made according to one of multiple coexisting choice criteria. We denote the set of possible choice criteria by $\mathcal{C}$. Thus the proposed model reduces to conventional Expected Utility theory when $|\mathcal{C}_{\text{EU}}| = 1$, whilst Dual-Process (thinking fast vs. thinking slow) decision making corresponds to a model with $|\mathcal{C}_{\text{DP}}| = 2$. We consider a more general case with $|\mathcal{C}| \geq 2$, which necessitates careful consideration of *how*, for a particular choice-task instance, one criterion comes to prevail over others. We operationalize this with a probability distribution that is conditional upon traits of the decision maker as well as upon the context and the framing of choice options. Whereas existing Signal Detection Theory (SDT) models of phishing detection commingle the different peripheral-route persuasion pathways, in the present descriptive generalization the different pathways are explicitly identified and represented. A number of implications follow immediately from this formulation, ranging from the conditional nature of security-breach risk to delineation of the prerequisites for valid tests of security training. Moreover, the model explains the 'stepping-stone' penetration pattern of APT attacks, which has confounded modeling approaches based on normative rationality.

Keywords: phishing, social engineering, peripheral-route persuasion, advanced persistent threat, choice criteria, dual-process theory, latent class model
JEL classification: D81, D91

†Financial support from the Economic and Social Research Council (UK) made this work possible, and is gratefully acknowle (grant number ES/P000665/1.).

1 INTRODUCTION

The human element in decision making is not only deliberative, but also emotional, intuitive, and fallible. Social-engineering campaigns target and exploit these non-deliberative features of human decision making.[1–7] A major lacuna for security-behavior modeling is that standard decision theory fails to capture the peripheral-route persuasion pathways that are exploited in social-engineering campaigns.

In contrast, Signal Detection Theory (SDT) has been successfully adapted to model human responses to phishing attacks.[1] The flexibility of SDT is instrumental in this context. This flexibility has been exploited to study the distinct consequences for security-breach risk estimates of premising the model solely upon normative decision theory, solely upon behavioral decision theory, or upon the combination of behavioral decision theory and susceptibility to peripheral-route persuasion.[8] Unsurprisingly, the latter combination proves most useful and informative. Nevertheless, two limitations may be observed in the existing SDT-based approach: (i) decision makers are assumed to be permanently characterized by one fixed decision-making model, and (ii) the effects of different peripheral-route persuasion pathways feed into, and become commingled in, a single value of the discriminability parameter.[2] Descriptive validity favors relaxation of the former, while interpretability of modeling favors relaxation of the latter.

We introduce a generalization of decision theory that fulfills these desiderata.[3] The generalization comprises two principal components.

First, a non-degenerate set $\mathcal{C}$ of 'ways of deciding' – here called 'choice criteria' – which in the phishing context includes not only Expected Utility (EU) to capture rational deliberative decision making, but also Prospect Theory (PT) which captures behavioral decision-making,[11] a 'routinely click-straight-through' element that captures unmotivated and unthinking routinized actions (automaticity),[12] and an 'impulsively click-through' element that captures emotionally motivated impulsive actions.[1–7] This approach therefore generalizes not only EU and PT, but also Dual-Process (DP) theories.[4]

[1] In a phishing attack, a network user receives an email containing either an attachment or a website link, which if opened, prompts the user to enter personal information (e.g. passwords) or infects the user's computer with malware that records such information surreptitiously. For SDT-based models of phishing vulnerability, see Kaivanto,[8] and Canfield and Fischhoff.[9]

[2] The standardized distance between the means of the sampling distributions, respectfully under the null and alternative hypotheses.

[3] The theory we present here is a specialization of Iain Embrey's 'States of Mind and States of Nature' formulation.[10]

[4] $\mathcal{C}_{\text{Here}} \supset \mathcal{C}_{\text{DP}} \supset \mathcal{C}_{\text{EU}}, \mathcal{C}_{\text{PT}}$.

It also formalizes the notion – to which the paper's title alludes – that there are several distinct types or classes of phishing ploy, and that individuals' susceptibility differs across qualitatively distinct social-engineering attack vectors. It is important to distinguish between these distinct phishing attack vectors – both to understand individuals' behavioral responses to them, and to understand organizations' total security-breach risk exposure. A phishing ploy that plays upon the prospect of a time-delimited opportunity for wealth is constructed very differently – and is processed very differently by its recipient(s) – than a phishing ploy that plays upon employees' standard routines of unquestioningly responding to bosses' and colleagues' emails, opening any appended email attachments, and clicking on enclosed links. An organization's email security training may effectively address the former, but in many organizations the latter remains a worrying vulnerability.

The second component of the generalization is a conditional probability distribution over the different choice criteria, i.e. over the elements of the set $\mathcal{C}$. As each new choice task is confronted, a draw from this distribution determines which choice criterion becomes operative, and so we will refer to it as the *State-of-Mind* (SoM) distribution for an individual i at time t. We allow an individual's SoM distribution to be conditional upon: their psychological traits and decision-experiences, the situational context of the decision, and the framing of the choice options. This approach is similar to that of two existing addiction models[13,14] although we extend those models by allowing the framing of the choice options to be strategically determined by an adversarial agent (the attacker), and by allowing both the prior-experiences and situational context of a decision to be strategically influenced by an allied agent (the Information Security Officer).

A key advantage of the present formulation is the top-level differentiation of the decision maker's susceptibility to different kinds of phishing ploys. This formulation yields a number of immediate implications. First, the overall security-breach risk due to phishing can not be conceived in unconditional terms. Since an individual's susceptibility to phishing depends on the type of phishing ploy, the phishing-ploy-*type* exposure distribution takes on importance, as does the intensity of this exposure (i.e. the total number of phishing emails traversing the spam filter) and the quality of phishing-ploy execution. Second, a single test-phishing email is insufficient for evaluating the effectiveness of email security training. Email security training does not necessarily generalize across different choice criteria. Hence, a single test-phishing email may determine

the robustness of security practice towards one particular phishing ploy, but it is orthogonal to potential vulnerabilities within the remaining choice criteria. Third, not only is the organization's security-breach risk conditional, but the attacker gets to *choose* the phishing-ploy-type exposure distribution, as well as the intensity of this exposure. The attacker has first-mover advantage. Moreover, the attacker always has the option to develop *new* phishing-ploy types that are not addressed by the organization's existing working practices and training materials. Fourth, given working practices in most organizations and given the dimensions over which the attacker can tailor a phishing campaign, it is clear that the attacker can attain a very high total probability of successfully breaching the target organization's cybersecurity. In part, this is due to the fact that typical working practices in non-high-security organizations[5] do not involve special treatment of embedded links or attached files.[6] It is also due to the *disjunctive* accumulation (addition, rather than multiplication) of successful-security-breach probabilities over spam-filter-traversing phishing emails. But it is also due to the scope for using rich contextual information to tailor a campaign into a *spear-phishing* attack – i.e. to specifically target the 'routinely click-straight-through' choice criterion characterized by automaticity.

Furthermore, our model supports an explanation for the 'stepping-stone penetration pattern' that is common in APT attacks. Whereas models of security behavior premised upon normative rationality have not been successful in explaining the stepping-stone pattern, we show that in light of a coexisting-choice-criteria model of security behavior, the stepping-stone penetration pattern may be recovered as a constrained-optimal attack vector.

The sequel is organized as follows. Section 2 briefly reviews the phishing literature, showing that phishing attacks employ social-engineering techniques that circumvent deliberatively rational decision processes. Section 3 reviews the empirical literature in which multiple 'ways of deciding' have been documented empirically, establishing a rigorous empirically grounded basis for the coexisting-choice-criteria model. Section 4 introduces the coexisting-choice-criteria model, and illustrates some of its properties, including its ability to support an explanation of the stepping-stone penetration pattern (Section 4.3). Section 5 concludes.

[5] e.g. commercial, administrative, professional-service, and higher-education organizations

[6] For instance the production of research articles involves multiple exchanges of emails among coauthors themselves, between the coauthors and the journal's editorial team, and then between the coauthors and the publisher's production team. These emails contain file attachments, and sometimes URLs as well. In these exchanges, there is no security procedure in place to authenticate emails, their attachments, or embedded URLs.

2 PHISHING TARGETS THE HUMAN ELEMENT

The capacity for rational deliberation is a feature of human beings, albeit not the overriding trait it was thought to be when Carl Linnaeus coined the binary nomenclature, *Homo sapiens*.[7] Both large-scale and narrowly targeted social engineering are predicated upon the intuitive, emotional, and fallible nature of human behavior, and it is now recognized that psychology is an essential component – alongside engineering and economics – for understanding information security.[15]

More than half of all US government network security-incident reports concern phishing attacks, and the number of phishing emails being sent to users of federal networks is growing rapidly.[16,17] The FBI and the DHS recently issued an amber alert warning of APT activity targeting energy – especially, nuclear power[8] – and other sectors.[19] In this broad APT campaign, spear phishing was the preferred initial-breach technique. The corporate sector is targeted more widely, commonly using phishing to create an entry point, for the purposes of extortion, illegally acquiring customer-information (and credentials) databases, as well as for acquiring commercially sensitive information. The incidence of corporate cyber espionage is not systematically disclosed, but many of the high-profile examples of corporate hacking that have come into the public domain were staged via phishing.[20]

Online scams such as phishing and spear phishing employ techniques of persuasion that have collectively been labeled 'social engineering'.[2,6] These techniques eschew direct, rational argumentation in favor of 'peripheral' routes to persuasion. The most prominent of these peripheral pathways to persuasion are, in no particular order: (i) authority, (ii) scarcity, (iii) similarity and identification, (iv) reciprocation, (v) consistency following commitment, and (vi) social proof.[1–7] Scams[9] typically augment peripheral-route persuasion by setting up a scenario that creates psychological pressure by triggering *visceral emotions* that override rational deliberation.[3,21,22] Visceral emotions – such as greed, envy, pity, lust, fear and anxiety – generate psychological discomfort as long as the underlying need remains unfulfilled, and psychological pleasure or even euphoria when that need is fulfilled. The manipulative scenario is deliberately structured so that the scammer's proposition offers the double prospect of relief from the visceral discomfort as

[7] Sapience denotes wisdom and the capacity for sound judgment, particularly in complex or difficult circumstances.

[8] For instance, the non-operational computer systems of Wolf Creek Nuclear Operating Corporation in Kansas were penetrated.[18]

[9] as well as 'hard-sell' and 'high-pressure' marketing more generally,

well as visceral satisfaction upon fulfilling the underlying need.

An ideally scripted scam scenario contrives a compelling, credible need for immediate action. If a scam-scenario script falls short of this ideal, it will almost invariably emphasize the urgency with which action must be taken.[3,21,22] In itself, this introduces visceral anxiety where none existed before, and simultaneously, precludes the availability of time for cooling off and for rational deliberation. Visceral emotions have both a direct hedonic impact as well as an impact via altering the relative desirability of different cues and attributes. Crucially, visceral emotions also affect how decision makers process information, narrowing and restricting attention to the focal hedonic cue and its availability (or absence) in the present.[21,22] Since visceral emotions – and their concomitant effects on attention and relative desirability of different cues/attributes – are short lived, scam scripts contrive reasons for immediate action.[10]

At sufficiently high levels of intensity, visceral emotions can override rational deliberation entirely.[21] Mass phishing scams often aim to exploit human emotions in this fashion. Spear phishing attacks, on the other hand, typically aim to exploit the intuitive and fallible nature of human decision making without necessarily stoking emotion. This approach targets the routinization and *automaticity*[12] upon which successful management of a high-volume inbox rests. For most civilian organizations outside the security community, employees trust emails – and any embedded URLs and file attachments – sent by bosses and immediate colleagues, and frequently also those sent by more distant contacts. Failure to do so would bring most organizations to a halt. Spear phishing thus exploits this routine and unquestioning trust that is automatically extended to bosses, colleagues, and contacts – and unintendedly, to plausible facsimiles thereof.

More surprising is the fact that spear phishing emails endowed with rich contextual information have been deployed successfully on both sides of the civilian/non-civilian and security/non-security divides. A partial list of successfully breached governmental, defense, corporate, and scientific organizations includes the White House, the Australian Government, the Reserve Bank of Australia, the Canadian Government, the Epsilon mailing list service, Gmail, Lockheed Martin, Oak Ridge National Laboratory, RSA SecureID, Coka Cola Co., Chesapeake Energy, and Wolf Creek Nuclear Operating Corporation.[16–18,20,24] When implemented well with appropriate contextual information, a spear-phishing email simply does not attract critical evaluation, and

[10] A former swindler relates the principle: "It is imperative that you work as quickly as possible. Never give a hot mooch time to cool off. You want to close him while he is still slobbering with greed."[23]

its contents are acted upon in a routine and automatic fashion.

3 COEXISTING CHOICE CRITERIA: EMPIRICAL PROVENANCE

Decision theorists are gradually coming to terms with the implications of dual-process theory, which has been developed by psychologists and recently popularized by Daniel Kahneman in *Thinking, Fast and Slow.*[25]

Meanwhile, a well-established stream of empirical-decision-theory literature offers legitimation for the notion that there may be more than one way of reaching a decision. That literature captures heterogeneity in choice criteria with Finite Mixture (FM) models. Standard estimation procedures for such models allow the data to determine how many different choice criteria are present, and then to provide, for each individual, the respective criterion-type membership probabilities. In Glenn Harrison and Elisabet Rutström's FM models,[11] the traditional single-criterion specification is statistically rejected, in their words providing "a decent funeral for the representative agent model that assumes only one type of decision process."[26] In turn, Coller et al.'s FM models show that "observed choices in discounting experiments are consistent with roughly one-half of the subjects using exponential discounting and one-half using quasi-hyperbolic discounting."[27] And using a Bayesian approach, Houser et al. show that different people use different decision rules – specifically, one of three different criteria – when solving dynamic decision problems.[28]

Multiple choice criteria are also well established in the empirical-game-theory literature. Stahl and Wilson fit an FM model to data on play in several classes of of 3×3 normal-form games, and find that players fall into five different boundedly rational choice-criteria classes.[29] Guessing games – a.k.a. Beauty-Contest games – have been pivotal in showing not only that backward induction and dominance-solvability break down, but also that game play can be characterized by membership in a boundedly rational, discrete (level-k) depth-of-reasoning class.[30] FM models are the technique of choice for analyzing Beauty-Contest data, revealing that virtually all 'non-theorist' subjects[12] (94%) fall into one of three boundedly rational depth-of-reasoning classes (levels 0, 1 or 2).[31,32] FM models are being applied increasingly in empirical game theory – including to the analysis of e.g. trust-game data, social-preferences data, and common-pool-

[11] of decision making under risk
[12] those who are not professional game theorists

resource data – demonstrating the broad applicability of a multiple-criteria approach. The theoretical relevance of level-k reasoning to adversarial interactions such as phishing has been further demonstrated by Rothschild et al.,[33] however we know of no existing paper in this field that allows alternative choice criteria to coexist.

Outside decision theory and empirical game theory, the necessity of allowing for multiple choice criteria has also been recognized in the fields of transportation research and consumer research. Within a Latent Class (LC) model framework,[13] Hess et al. study the question of whether "actual behavioral processes used in making a choice may in fact vary across respondents within a single dataset."[34] Preference heterogeneity documented in conventional single-choice-criterion models[14] may be a logical consequence of the single-choice-criterion restriction (i.e. misspecification). Hess et al. account for choice-criterion heterogeneity in four different transport-mode-choice datasets by fitting LC models. These LC models distinguish between conventional random utility and the lexicographic choice criterion (dataset 1), among choice criteria with different reference points (dataset 2),[15] between standard random utility and the elimination-by-aspects choice criterion (dataset 3), and between standard random utility and the random-regret choice criterion (dataset 4).[34] Finally, Swait and Adamowicz show that *consumers* also fall into different 'decision strategy' LCs, and that increasing either the complexity of the choice task or the cumulative task burden induces switching toward simpler decision strategies.[35] These results underscore an interpretation of the choice-criterion probabilities that is only implicit in the above-mentioned studies: that (a) decision makers should not be characterized solely in terms of their *modal* choice criterion, but in terms of their choice-criterion mixtures, and that (b) the criterion that is operative for a particular choice task is obtained as a draw from the probability distribution over choice criteria, which in turn is conditional upon features of the context, the framing and presentation of the choice options,and the current psychological characteristics of the decision maker.

In light of these FM- and LC-model findings, accommodation of multiple choice criteria emerges as a natural step toward improving the descriptive validity of theoretical models.

[13] Latent Class (LC) models are specializations of FM models.
[14] e.g. heterogeneity in risk aversion in EU models, and heterogeneity in probability weighting in PT models
[15] note that standard random utility has no reference point

4 INCORPORATING *Homo intuivus, emotus et fallibilis*

4.1 Coexisting-choice-criteria model

The econometric evidence reviewed in Section 3 warrants a generalization of decision theory to incorporate multiple coexisting choice criteria. An abstract formulation of such a theory naturally draws upon the formal specification of econometric latent class models that capture choice-criterion heterogeneity.[34,35]

Let $\mathcal{C}$ denote the set of coexisting choice criteria. The elements of this set are distinguished by the integer-valued index c, where $1 \leq c \leq C := |\mathcal{C}|$.

We specialize the present formulation to the context of phishing-security modeling by populating the set of choice criteria $\mathcal{C}$ with a view to capturing the essential features of human beings in the security setting, as reviewed in Section 2. Email recipients are capable of rational deliberation, but they are not overwhelmingly predisposed to it. They may instead form subjective beliefs and valuations as captured by behavioral decision theory, but they also frequently act in an intuitive or routinized fashion. Thus the empirical evidence reviewed in Section 2 suggests that human responses to phishing campaigns range across (at least) four identifiable choice criteria, which we summarize in Table 1.[16]

Table 1: Email recipients' coexisting choice criteria.

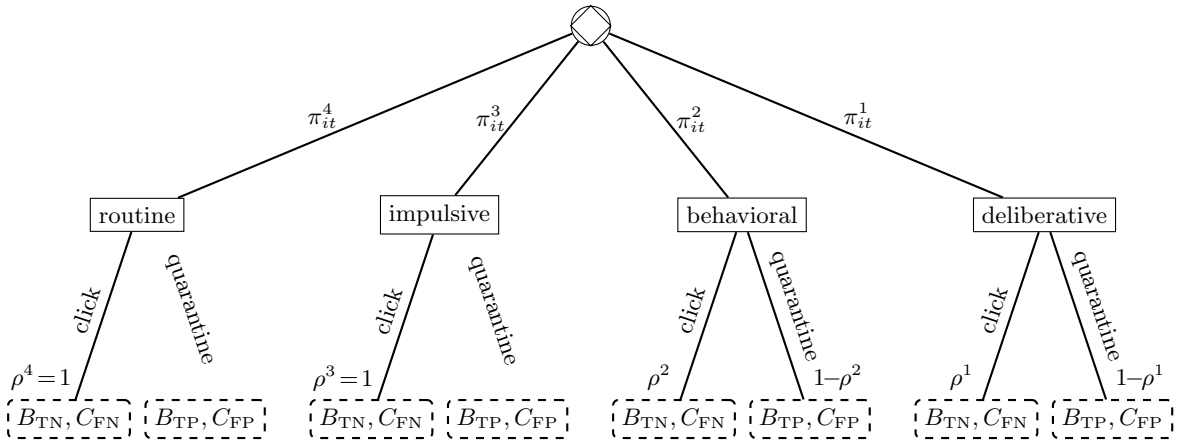
$c=1$	Normative deliberation: characterized by the internal-consistency axioms of completeness, transitivity, independence of irrelevant alternatives (iia), continuity, Bayesian updating, and time consistency (i.e. exponential discounting).
$c=2$	Behavioral: characterized by the weakening of iia, Bayesian updating, and time consistency (i.e. to hyperbolic discounting), as per the behavioral decision making literature.
$c=3$	Impulsively click through: characterized by dominance of visceral emotions, which suppress and displace deliberative reasoning; the remaining consistency axioms are abandoned.
$c=4$	Routinely click straight through: characterized by routinization and automaticity; again, the remaining consistency axioms are abandoned.

[16] Although the present formulation moves beyond the simplicity of a single-decision-criterion world view, each criterion of Table 1 can be formalized by an existing theoretical framework. Normative deliberation is underpinned by the axiomatizations of e.g. von Neumann and Morgenstern, or Leonard Savage. Descriptively valid, partly deliberative behavioral rationality is underpinned by axiomatizations of Cumulative Prospect Theory by e.g. Wakker and Tversky.[36] The deliberative-rationality-displacing role of visceral emotions has been recognized in the evolutionary study of behavior, represented in economics in particular by e.g. Robert H. Frank.[37] Automaticity, in which deliberative rationality is not so much bypassed as simply 'not engaged', has been given theoretical underpinning in the psychology literature by Moors and De Houwer.[12] Imperfect and fallible recognition, categorization, and procedural responses have been widely documented,[38–42] for which general theoretical underpinning may be obtained from e.g. Philippe Jehiel's concept of analogy-based expectations equilibrium.[43]

In general the choice-criterion selection probability will be conditional upon the decision maker's State of Mind, which in turn depends on an array of subject- and task-specific variables. The net effect of all such variables determines an individual's probability of adopting a given choice criterion c at a given point in time, which we denote by π_{it}^{c}. Note that we necessarily have $0 \leq \pi_{\text{it}}^{c} \leq 1$ and $\sum_{c=1}^{C} \pi_{\text{it}}^{c} = 1$ for all individuals i and time-points t.

Figure 1 illustrates a single agent's stochastic State-of-Mind response to an arbitrary email. This begins with the diamond-within-a-circle chance node, whereby the incoming email probabilistically triggers one of the four State-of-Mind choice criteria. The fact that the 'Routine' (c=4) and 'Impulsive' (c=3) choice criteria override the possibility of sufficient deliberation to result in a 'quarantine' choice with probability $\rho = 1$ is indicated by the absence of these respective edges. The email recipient's incomplete information – over whether the email is benign or malicious – is reflected in the broken-line information sets surrounding terminal-node payoffs.

Figure 1: An agent's stochastic State-of-Mind response to an email.



Note: Ex ante the agent is uncertain about an email's true nature. The payoff at each terminal node is therefore either a benefit due to correct classification (True Positive or True Negative), or a cost due to incorrect classification (FP or FN).

The email recipient is one of many agents who interact in a strategic phishing game. We analyze an attacker's optimal response to Figure 1 in Section 4.3, and we discuss the model's implications for organizational security policy in Section 4.4. Before doing so, we complete the model by expanding the π_{it}^{c} expression for an agent i at time t. In general, π_{it}^{c} is operationalized through a probability distribution that may be conditional upon: the characteristics of the

decision maker X_{it}, the situational context Z_{it}, and the attributes of the present choice task $\boldsymbol{\alpha}_t$.

$$\pi_{it}^c = \pi^c\big(X_{it}, Z_{it}, \boldsymbol{\alpha}_t\big), \qquad 0 \leq \pi_{it}^c \leq 1, \qquad \sum_{c=1}^{C} \pi_{it}^c = 1, \tag{4.1a}$$

$$X_{it} = f\Big(\Gamma_i,\ \{Z_i\}_{<t}\,,\ \{\boldsymbol{\alpha}\}_{<t}\,,\ \{D_i\}_{<t}\Big)\ . \tag{4.1b}$$

The current characteristics X_{it} of agent i are jointly determined by their stable psychological traits Γ_i, and by the history of: decision contexts $\{Z_i\}_{<t}$, decision-attributes $\{\boldsymbol{\alpha}\}_{<t}$, and decision-outcomes $\{D_i\}_{<t}$ that constitutes their current set of experiences.

In order to develop a tractable expression for π_{it}^c we generalize the notion of match quality introduced in the SDT literature[8] and we specialize the vectors appearing in (4.1a) to the phishing-email application. For this application, the context Z_{it} is that in which the agent receives his emails. An agent whose context Z_{it} and recent context history $\{Z_i\}_{<t}$ leaves him stressed, distracted, or hungry, will be less likely to respond deliberatively. The implications of this observation for personal practice and organizational security policy are clear,[17] and so we suppress Z_{it} hereafter to focus on the strategic interaction between attackers and recipients. For simplicity we also suppress time subscripts hereafter to focus on the short-run implications of the model.

Let us consider a phishing email with attributes $\boldsymbol{\alpha}$ constructed within a finite attribute space $\mathcal{A} = [0,1]^A$. Each of the A components of email $\boldsymbol{\alpha}$ captures the emphasis that it places on each of A possible cues. The attacker chooses which cues to emphasize in order to influence the recipient's State of Mind. This determination of email 'content' is the attacker's primary decision variable.

The attacker is nevertheless constrained, in that increasing the emphasis placed on any one cue necessarily diminishes the emphasis on the others. We model this constraint by requiring $||\boldsymbol{\alpha}|| \leq 1$.

The salient characteristics of the recipient are his idiosyncratic susceptibility to each type of cue S_i, and his baseline propensity χ_i^c to apply each choice criterion c.[18] S_i is an $C \times A$ dimensional matrix, each row of which $\boldsymbol{s}_i^c$ specifies the effectiveness of each possible cue type

[17] These are discussed further in Section 4.4,

[18] The baseline propensity to adopt a deliberative choice criterion χ_i^1 is a stable trait[44] that can be measured by the cognitive Reflection Test of Frederick[42], or the Decision-Making Competence scale of Parker and Fischhoff[45].

in invoking the choice criterion c. The agent's characteristics X_i are therefore a matrix in $[0,1]^{A\times C} \times (\mathbb{R}^+)^C$, each row of which is a pair $\{\boldsymbol{s}_i^c, \chi_i^c\}$ that will determine the match quality between the attacker's choice of email cues $\boldsymbol{\alpha}$, and the susceptabilities of the receiving agent i.

We may now extend the approach of Kaivanto[8] by defining the choice-criterion-specific match-quality function $m^c : [0,1]^A \times [0,1]^A \times \mathbb{R}^+ \to \mathbb{R}^+$, such that

$$m_i^c(\boldsymbol{\alpha}) = m^c(\boldsymbol{\alpha}, \boldsymbol{s}_i^c, \chi_i^c) \qquad \forall\, c \in \mathcal{C} \ . \tag{4.2}$$

For illustrative purposes, the simplest non-degenerate functional form for m^c would be the separable linear specification

$$m_i^c(\boldsymbol{\alpha}) = \chi_i^c + \boldsymbol{s}_i^c \cdot \boldsymbol{\alpha} \ , \tag{4.3}$$

where $\cdot$ denotes the vector dot product.

Agent i's choice-criterion-selection probabilies for a given email with cue bundle $\boldsymbol{\alpha}$ may then be defined in terms of the match-quality functions as follows:

$$\pi_i^c(\boldsymbol{\alpha}) = \frac{m_i^c(\boldsymbol{\alpha})}{\sum_{c\in\mathcal{C}} m_i^c(\boldsymbol{\alpha})} \qquad \forall\, c \in \mathcal{C} \ . \tag{4.4}$$

4.2 Contrast with normatively rational deliberative special case

Under a normative decision-theoretic model of email-recipient decision making it is difficult to explain the existence of phishing as an empirical phenomenon. Normatively rational decision making is a special case of the coexisting-choice-criteria model in which $\pi^1 = 1$ and $\pi^2 = \pi^3 = \pi^4 = 0$. If all email recipients were characterized by choice-criterion #1 alone, then the success of an email phishing campaign would be determined entirely by factors largely outside the attacker's control: the benefit from correctly opening a non-malicious email (B_{TN}), the cost of erroneously quarantining non-malicious email (C_{FP}), the cost of erroneously opening a malicious email (C_{FN}), and the benefit of correctly quarantining a malicious email (B_{TP}). Instead, variation in phishing campaigns' success rate is driven by factors that do not directly affect $B_{\text{TN}}, C_{\text{FP}}, C_{\text{FN}}$ and B_{TP}.[2,4,6,7]

It is straightforward to explain the existence of phishing and its empirical characteristics under a coexisting-choice-criteria model of email-recipient behavior in which $\pi^1 < 1$ and $\pi^2, \pi^3, \pi^4 > 0$. For instance choice-criterion #4 (routine, automaticity) is triggered by a phish-

ing email that masquerades as being part of the normal work flow by exploiting rich contextual information about the employee, the organizational structure (e.g. boss' and colleagues' names, responsibities, and working practices), and current organizational events and processes. Here the email recipient simply does not engage in a deliberative process to evaluate whether the email should be opened or not.

In contrast, phishing ploys designed to trigger choice criterion #3 (impulsively click through) employ what Robert Cialdini calls the *psychological principles of influence* (see Section 2).[2,4–7] Importantly, there is variation between individuals in their susceptibility to particular levers of psychological influence.[7,46,47] For instance scarcity[19] and authority[20] have been found to be more effective for young users, while reciprocation[21] and liking/affinity[22] have been found to be more effective for older users.[7] These observations motivate the agent-specific subscript i in π_i^c and m_i^c, and they are important in establishing the constrained-optimal APT attack pattern in the following Subsection.

None of the aforementioned psychological levers would be effective if email users were solely $c \equiv 1$ normatively rational deliberators. Similarly, the well-documented effects of commitment,[23] perceptual contrast,[24] and social proof [25] (see[2,4,6,7]) are naturally explained by the existence of coexisting choice criteria.

4.3 Stepping-stone penetration

Forensic investigations of APT attacks have found that the initial breach point is typically several steps removed from the ultimate information-resource target(s). Deliberation-based models of normatively rational decision making offer no particular insight into this empirical regularity. In contrast, the coexisting-choice-criteria model encodes differentiation with which the stepping-stone penetration pattern may be recovered as a constrained-optimal attack vector.

Let us consider an attacker who wishes to achieve a click-through from one of a minority

[19] e.g. Don't miss out on this 'once-in-a-lifetime opportunity!'

[20] e.g. law enforcement officers, tax officials

[21] the tendency to repay in kind even though there is no implicit obligation to do so

[22] the tendency to comply with requests made by people whom the user likes or with whom the user shares common interests or common affiliations

[23] Also referred to as 'consistency'. People feel obliged to behave in line with – consistently with – their previous actions and commitments.

[24] Making an option seem attractive by framing it with respect to an option that is (contrived to be) noticeably less attractive.

[25] People conform with majority social opinion, even when this manifestly contradicts immediate personal perception, as in e.g. the Stanford Prison Experiment.

subset of m target individuals within an organization consisting of n members. The target individuals may be those who can authorize expenditure, or those with particular (e.g. database) access rights. The attacker's strategy at any given point in time consists of a choice of cue-bundle $\boldsymbol{\alpha}_k$, taken to solve the program

$$\max_{\boldsymbol{\alpha}_k} \quad \sum_{i=1}^{m} \sum_{c=1}^{C} \pi_i^c(\boldsymbol{\alpha}_k) \cdot \rho_i^c \cdot V \quad - \; e(\boldsymbol{\alpha}_k) \qquad \text{s.t.} \qquad ||\boldsymbol{\alpha}_k|| \leq 1 \;\; , \tag{4.5}$$

where $\pi_i^c(\boldsymbol{\alpha}_k)$ is the probability with which an individual will adopt choice criterion c given the cues present in phishing email $\boldsymbol{\alpha}_k$, where ρ^c is the probability of click-through given choice criterion c, where V is the expected value of a successful attack, and where $e(\boldsymbol{\alpha}_k)$ is the cost of the effort expended in the production and distribution of email $\boldsymbol{\alpha}_k$. This formulation accords with the near-zero marginal cost of including additional recipients to any existing email.[48,49]

The attacker may send one, or more, emails $\boldsymbol{\alpha}_k$. Each email may be designed to induce one particular State-of-Mind c, or could in principle adopt a mixed strategy. However, since (by construction and by necessity) $\sum_{c \in \mathcal{C}} \pi_i^c = 1$, any mixture of asymmetrically effective pure strategies must be strictly less effective than at least one pure strategy. We therefore proceed by characterizing the available pure strategies on the basis of the phishing literature,[2,4,6] before eliminating strictly dominated strategies.

Table 2: Choice-criterion targeting characteristics.

Choice criterion	Effort	Click-through prob.	Selection prob.[a]	
c	$e\big(\underset{\boldsymbol{\alpha}}{\text{argmax}}\{\pi^c\}\big)$	ρ^c	Prior	Posterior
$c=1$: Deliberative	low	negligible	high	high
$c=2$: Behavioral	low	low	med	med
$c=3$: Impulsive	low	1	low	low
$c=4$: Routine	high	1	low	high

[a] i.e. $\max_{\boldsymbol{\alpha}}\{\pi^c\}$

The quantities summarized in Table 2 determine the costs and expected benefits to the attackers of targeting choice criterion c through their choice of $\boldsymbol{\alpha}$. There are two values of the selection probability $\max_{\boldsymbol{\alpha}}\{\pi^c\}$ for each choice criterion c: the prior likelihood of invoking that criterion, without insider information, and the posterior likelihood once access to such insider information is obtained. Insider information does not affect the attackers' ability to invoke choice

criteria $c \in \{1, 2, 3\}$, but it does greatly aid the attacker's ability to 'spoof' (i.e. simulate) a routine email from a trusted colleague, and hence it substantially increases the posterior selection probability for $c = 4$. The mechanism by which attackers may gain such insider information is the successful phishing of a non-target member of the organization.

The most immediate implication of Equation (4.5) and Table 2 is that the Deliberative strategy is strictly dominated by the Behavioral strategy, due to the negligible click-through probability of the former. We next observe that the Behavioral strategy is, in turn, strictly dominated by the Impulsive strategy whenever

$$\rho^2 < \frac{\max_{\boldsymbol{\alpha}}\{\pi^3\}}{\max_{\boldsymbol{\alpha}}\{\pi^2\}} \ , \tag{4.6}$$

That is whenever the expected click-through probability under a Behavioral choice criterion is less than the relative ease of invoking the Behavioral state compared to invoking the Impulsive state. Table 2 suggests that this criterion is typically satisfied.

Next we consider the case of an attacker who has no insider information. In this case it is trivial to see that an email which aims to invoke the Impulsive choice criterion strictly dominates an email which aims to invoke the Routine choice criterion, due to the lower effort cost of the former. The respective probabilities of successfully gaining a click-through from a target individual are then:

$$\text{Prob.}\left(\begin{smallmatrix}\text{non-target}\\ \text{clickthrough}\end{smallmatrix}\right) = 1\text{–}(1\text{–}\max_{\boldsymbol{\alpha}}\{\pi^3\})^{n-m} \quad > \quad 1\text{–}(1\text{–}\max_{\boldsymbol{\alpha}}\{\pi^3\})^{m} = \text{Prob.}\left(\begin{smallmatrix}\text{target}\\ \text{clickthrough}\end{smallmatrix}\right)$$

which demonstrates that there is a greater likelihood of the attacker gaining a click-through from a non-target individual than from a target individual in any attack without insider information. Note that this conclusion would be further strengthened if we were to assume that target individuals were less susceptible to phishing attacks than the average individual.

The attackers' first attempt therefore has three possible outcomes: (i) they may have successfully achieved their objective, (ii) they may have gained insider information by achieving a non-target click-through, or (iii) they may have achieved nothing. In the first case the attackers move on to acquire and exfiltrate the information. In the third case the situation is unchanged, and so the phishing campaign is continued with further broadcast of phishing email(s) containing (possibly modified) Impulsive cues. But in the second case insider information is obtained,

whereby the posterior click-through likelihoods of Table 2 become operative. In this case, it is evident from Table 2 that an email which aims to invoke the Routine choice criterion is likely to dominate an email which aims to invoke the Impulsive criterion, specifically whenever

$$\frac{e\big(\text{argmax}_{\boldsymbol{\alpha}}\{\pi^4\}\big)}{e\big(\text{argmax}_{\boldsymbol{\alpha}}\{\pi^3\}\big)} < \frac{\max_{\boldsymbol{\alpha}}\{\pi^4\}}{\max_{\boldsymbol{\alpha}}\{\pi^3\}} \quad . \tag{4.7}$$

Thus the attacker's optimal approach is likely to lead to a 'stepping-stone' attack, wherein a non-target individual is first compromised by invoking an impulsive choice criterion, so that a target individual can then be compromised by using insider information to invoke a Routine choice criterion. Sufficient conditions for this to be the most likely outcome are those of Table 2 and inequalities (4.6) and (4.7).

4.4 Implications for Organizational Security Policy

The model we present has important implications for organizational security policy. Let us first consider the cultural and procedural aspects of organizational security, before turning to specific implications for email security training and evaluation.

In Section 4.1 we noted the potential importance of the situational context Z_{it} in which an email is received. For example, it is well-known that an individual who is under intense time-pressure is less likely, if not not simply unable, to engage in deliberative decision making.[50–52] The present model makes plain the security-vulnerability dangers of highly routinized email-processing practices, even if these would otherwise be efficient. Relatedly, it is vital that organizational culture supports the precautionary verification of suspicious messages, since any criticism of such verification practices is likely to increase the risk of behavioral click-throughs in future. These observations suggest that Information Security Officers should actively engage with wider aspects of organizational culture and practices.

The model also yields specific procedural implications for email security training. It is clear that the direct effect of a training course in which participants consciously classify emails as either genuine or malicious would be to reduce ρ^1 (see Figure 1), however for most individuals ρ^1 is already relatively low (see Table 2): given that an individual implements a deliberative choice criterion they are relatively unlikely to fall prey to a phishing attack. Section 4.3 demonstrated that a strategic attacker would instead seek to exploit the much greater vulnerabilities of ρ^3 and

ρ^4, and so training that focuses on reducing ρ^1 is likely to have limited effectiveness.

The challenge for Information Security Officers is that the vulnerabilities ρ^3 and ρ^4 are essentially fixed at 1.[26] Once an Impulsive or Routine State of Mind takes over, click-through is a foregone conclusion. Training should therefore focus on reducing individuals' criterion-selection probabilities π^3 and π^4. There is evidence that an individual's propensity to act deliberatively can be raised through external interventions,[42] and the coexisting-choice-criteria framework suggests that this could best be achieved by helping employees to understand: (i) their inherent vulnerability to phishing when making choices either Routinely or Impulsively, and (ii) the psychological ploys by which attackers may induce Impulsive or Routine States of Mind.

Analogous implications exist for procedures which aim to test organizational security by means of simulated phishing emails. Where such a test is appended to a training module, it tests (at best) some combination of ρ^1 and ρ^2, because trainees will be aware that they are attempting to identify phishing emails. Furthermore, the literature on incentives suggests that where such a test is incentivized with some required pass-rate, it is likely to be less informative as to the true vulnerability level because it is more likely to generate a pure measure of ρ^1. Tests of security should therefore be blinded, for example by an unannounced simulation of an email attack. Moreover, such tests should be varied and repeated, since any single email $\boldsymbol{\alpha}$ can only contain one specific cue bundle, and so can only test an individual's susceptibility $\pi^c(\boldsymbol{\alpha})$ to that particular cue bundle.

5 CONCLUSION

As the basis for understanding and modeling the behavior of phishing targets, normative deliberative rationality proves wholly inadequate. This paper introduces a coexisting-choice-criteria model of decision making that generalizes both normative and 'dual process' theories of decision making. We show that this model offers a tractable working framework within which to develop an understanding of phishing-email response behavior. This offers an improvement over existing SDT-based models of phishing-response behavior,[8,9] insofar as it avoids the commingling of peripheral-route-persuasion pathways.

We also show that the framework may be usefully deployed in modeling the choices and

[26] In principle an organization could substantially reduce ρ^4 by implementing organization-wide 2-stage security procedures before any email link or attachment is opened, however such measures have not been widely adopted due to their efficiency cost (as discussed in Sections 1 and 2).

tradeoffs confronted by APT attackers, who must make decisions about the nature, composition, and roll-out of phishing campaigns. We illustrate this by tackling a problem that has confounded conventional normative-rationality-based modeling approaches: Why do so many APT attacks follow a 'stepping-stone' penetration pattern? Under the coexisting-choice-criteria model, the attacker faces a tradeoff between (i) designing an email that is highly targeted, invokes the 'Routine' choice criterion, but requires detailed inside information, and (ii) designing an email that cannot be targeted as effectively, invokes the 'Impulsive' choice criterion, and requires only public information. However, success with (ii) provides the attacker with access to the inside information with which to implement (i). Thus the stepping-stone attack vector arises out of the attacker's tradeoffs precisely when confronting email users whose behavior is captured by the coexisting-choice-criteria model.

We further demonstrate that the model provides new insights with practical relevance for Information Security Officers. We derive specific recommendations for information training and testing as well as for organizational procedures, practices, and policies. In particular, the model highlights the importance of considering the composite between the probability of being induced into State of Mind c and the probability of then clicking through *given* this State of Mind. Hence training must address the different State-of-Mind selection probabilities π^c as well as the associated conditional click-through probabilities ρ^c. Similarly, training-effectiveness testing – assurance of learning, in effect – must also cover the range of different State-of-Mind choice criteria. In light of the coexisting-choice-criteria model, the single-test-email approach should be deprecated.

Finally, the coexisting-choice-criteria model highlights organizations' vulnerability to spear-phishing attacks that invoke automatic email processing routines. Working practices in most commercial, voluntary, and public-sector organizations presume that links and email attachments are benign when sent from within the organization or by customers, suppliers, or partner organizations. This is a major vulnerability that is as much a reflection of organizational culture as it is a reflection of explicit security protocols (or absence thereof). This suggests that Information Security Officers could – and perhaps should – be afforded a broader role in shaping organizational culture.

References

1. Petty RE, Cacioppo JT. Communication and Persuasion: Central and Peripheral Routes to Attitude Change. New York, NY: Springer-Verlag, 1986.

2. Rusch JJ. The "social engineering" of internet fraud. Proceedings of the Internet Society Global Summit (INET'99), 1999, June 22–25, San Jose, CA. http://www.isoc.org/inet99/proceedings/3g/3g_2.htm

3. Langenderfer J, Shimp TA. Consumer vulnerability to scams, swindles, and fraud: A new theory of visceral influences on persuasion. Psychology and Marketing, 2001; 18:763–783.

4. Mitnick KD, Simon WL. The Art of Deception: Controlling the Human Element of Security. Indianapolis, IN: Wiley, 2002.

5. Cialdini RB. Influence: The Psychology of Persuasion. New York, NY: Collins, 2007.

6. Hadnagy C. Social Engineering: The Art of Human Hacking. Indianapolis, IN: Wiley, 2011.

7. Oliveira D, Rocha H, Yang H, Ellis D, Dommaraju S, Muradoglu M, Weir D, Soliman A, Lin T, Ebner N. Dissecting spear phishing emails for older vs young adults: On the interplay of weapons of influence and life domains in predicting susceptibility to phishing. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 6412–6424. http://chi2017.ac.org/proceedings.html

8. Kaivanto K. The effect of decentralized behavioral decision making on system-level risk. Risk Analysis, 2014; 34:2121–2142.

9. Canfield CI, Fischhoff B. Setting priorities for behavioral interventions: An application to reducing phishing risk. Risk Analysis, in press. doi: 10.1111/risa.12917.

10. Embrey I. States of nature and states of mind: A generalised theory of decision-making, evaluated by application to human capital development. Working Paper, Lancaster University Management School, Economics Working Paper Series, 2017; 2017/032. http://eprints.lancs.ac.uk/89106/1/LancasterWP2017_032.pdf

11. Tversky A, Kahneman D. Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty, 1992; 5:297–323.

12. Moors A, DeHouwer J. Automaticity: A theoretical and conceptual analysis. Psychological Bulletin, 2006, 132:297–326.

13. Laibson D. A cue-theory of consumption. The Quarterly Journal of Economics, 2001; 116:81–119.

14. Bernheim D, Rangel A. Addiction and cue-triggered decision processes. The American Economic Review, 2004; 94:1558–1590.

15. Anderson R, Moore T. Information security: Where computer science, economics and psychology meet. Philosophical Transactions of the Royal Society A, 2009; 367;2717–2727.

16. US Office of Management and Budget. Fiscal Year 2011 Report to Congress on the Implementation of The Federal Information Security Management Act of 2002. March 7, 2012.

17. Johnson NB. Feds' chief cyberthreat: 'Spear phishing' attacks. Federal Times, Feb 20, 2013.

18. Perlroth N. Hackers are targeting nuclear facilities, Homeland Security Dept. and F.B.I. say. New York Times, July 6, 2017.

19. FBI and DHS. Advanced Persistent Threat Activity Targeting Energy and Other Critical Infrastructure Sectors. 'Amber' Alert (TA17-293A), Oct 20, 2017. https://www.us-cert.gov/ncas/alerts/TA17-293A

20. Elgin B, Lawrence D, Riley M. Coke gets hacked and doesn't tell anyone. Bloomberg, Nov 4, 2012. http://www.bloomberg.com/news/2012-11-04/coke-hacked-and-doesn-t-tell.html

21. Loewenstein G. Out of control: Visceral influences on economic behavior. Organizational Behavior and Human Performance, 1996; 65:272–292.

22. Loewenstein G. Emotions in economic theory and economic behavior. American Economic Review, 2000; 90:426–432.

23. Easley B. Biz-Op: How to Get Rich with "Business Opportunity" Frauds and Scams. Port Townsend, WA: Loompanics Unlimited, 1994.

24. Hong J. The state of phishing attacks. Communications of the ACM, 2012; 55:74–81.

25. Kahneman D. Thinking, Fast and Slow. New York, NY: Penguin, 2012.

26. Harrison GW, Rutström EE. Expected utility theory and prospect theory: One wedding and a decent funeral. Experimental Economics, 2009; 12:133–158.

27. Coller M, Harrison GW, Rutström EE. Latent process heterogeneity in discounting behavior. Oxford Economic Papers, 2011; 64:375–391.

28. Houser D, Keane M, McCabe K. Behavior in a dynamic decision problem: An analysis of experimental evidence using a Bayesian type classification algorithm. Econometrica, 2004; 72:781–822.

29. Stahl DO, Wilson PW. On players' models of other players: Theory and experimental evidence. Games and Economic Behavior, 1995; 10:218–254.

30. Nagel R. Unraveling guessing games: An experimental study. American Economic Review, 1995; 85:1313–1326.

31. Stahl DO. Boundedly rational rule learning in a guessing game. Games and Economic Behavior, 1995; 16:303–313.

32. Bosch-Domènech A, Montalvo JG, Nagel R, Satorra A. A finite mixture analysis of beauty-contest data using generalized beta distributions. Experimental Economics, 2010; 13:461–475.

33. Rothschild C, McLay L, Guikema S. Adversarial risk analysis with incomplete information: A level-k approach. Risk Analysis, 2012; 32:1219–1231.

34. Hess S, Stathopoulos A, Daly A. Allowing for heterogeneous decision rules in discrete choice models: An approach and four case studies. Transportation, 2012; 39:565–591.

35. Swait J, Adamowicz W. The influence of task complexity on consumer choice: A latent class model of decision strategy switching. Journal of Consumer Research, 2001; 28:135–148.

36. Wakker P, Tversky A. An axiomatization of cumulative prospect theory. Journal of Risk and Uncertainty, 1993; 7:147–175.

37. Frank, RH. Passions Within Reason: The Strategic Role of the Emotions. New York, NY: Norton, 1988.

38. Chou E, McConnell M, Nagel R, Plott CR. The control of game form recognition in experiments: Understanding dominant strategy failures in a simple twoperson guessing game. Experimental Economics, 2009; 12:159–179.

39. Kaivanto K, Kroll EB, Zabinski M. Bias-trigger manipulation and task-form understanding in Monty Hall. Economics Bulletin, 2014; 34:89–98.

40. Goldstein DG, Taleb NN. We don't quite know what we are talking about when we talk about volatility. Journal of Portfolio Management, 2007; 22:84–86.

41. VanLehn K. Mind bugs: The Origins of Procedural Misconceptions. Cambridge, MA: MIT Press, 1990.

42. Frederick S. Cognitive reflection and decision making. Journal of Economic Perspectives, 2005; 19:25–42.

43. Jehiel P. Analogy-based expectation equilibrium. Journal of Economic Theory, 2005; 123:81–104.

44. Parker AM, De Bruin WB, Fischhoff B, Weller J. Robustness of decision-making competence: Evidence from two measures and an 11-year longitudinal study. Journal of Behavioral Decision Making, in press. doi:10.1002/bdm.2059

45. Parker AM, Fischhoff B. Decision-making competence: External validation through an individual-differences approach. Journal of Behavioral Decision Making, 2005; 18;1–27.

46. Vishwanath A, Herath T, Chen R, Want J, Rao HR. Why do people get phished? Testing indiviual differences in phishing vulnerability within an integrated, information processing model. Decision Support Systems, 2011; 51:576–586.

47. Williams EJ, Beardmore A, Joinson AJ. Individual differences in susceptibility to online influence: A theoretical review., Computers in Human Behavior, 2017; 72:412–421.

48. Shapiro C, Varian HR. Information Rules: A Strategic Guide to the Network Economy. Boston, MA: Harvard Business School Press, 1998.

49. Anderson RJ. Security Engineering: A Guide to Dependable Distributed Systems. Indianapolis, IN: John Wiley, 2008.

50. Hwang MI. Decision making under time pressure: A model for information systems research. Information & Management, 1994; 27:197–203.

51. Maule AJ, Edland AC. The effects of time-pressure on judgment and decision making. In: Ranyard R, Crozier W, Svenson O (eds). Decision Making: Cognitive Models and Explanation. London: Routledg, 1997:189–204.

52. Steigenberger N, Lübcke T, Fiala H, Riebschläger. Decision Modes in Complex Environments. London: CRC Press (Taylor & Francis), 2017.