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

Iain Embrey and Kim Kaivanto

Lancaster University, Lancaster LA1 4YX, UK

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 $|\mathcal{C}_{\mathrm{EU}}| = 1$, whilst Dual-Process (thinking fast vs. thinking slow) decision making corresponds to a model with $|\mathcal{C}_{\mathrm{DP}}| = 2$. 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, Advanced persistent threat, Social engineering

1 Introduction

The human element in decision making is not only deliberative, but also emotional, intuitive, and fallible. Social-engineering attacks target and exploit these non-deliberative features of human decision making.[1–6] A major lacuna for security-behavior modeling is that standard decision models fail to capture the peripheral-route persuasion pathways exploited by social-engineering attacks.

In contrast, Signal Detection Theory (SDT) has been successfully adapted to model human responses to phishing attacks.[7, 8] 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.[7] 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.[1] Descriptive validity favors relaxation of the former; interpretability of modeling favors relaxation of the latter.

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

First, a non-degenerate set $\mathcal{C}$ of 'ways of deciding' – here called 'choice criteria'. In the phishing context this includes not only Expected Utility (EU) to capture rational deliberative decision making, but also: Prospect Theory (PT) capturing behavioral decision-making,[10] a 'routinely click-straight-through' element that captures unmotivated and unthinking routinized actions (automaticity),[11] and an 'impulsively click-through' element that captures emotionally motivated impulsive actions.[1–6] Our approach therefore generalizes not only EU and PT, but also Dual-Process (DP) theories ($\mathcal{C}_{\text{Here}} \supset \mathcal{C}_{\text{DP}} \supset \mathcal{C}_{\text{EU}}, \mathcal{C}_{\text{PT}}$).

The set $\mathcal{C}$ 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 pitches a time-limited opportunity for wealth is constructed very differently – and is processed very differently by its recipient(s) – to a phishing ploy that plays upon employees' standard routines of unquestioningly responding to their bosses' 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.

Second, 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. We refer to it as the *State-of-Mind* (SoM) distribution for an individual i at time t.[2] 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. Thus any target individual on any given occasion *may* deliberate rationally as to whether the present email is malicious, *or* their action may be determined by behavioural or automatic choice criteria.

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 for-

[1] The standardized distance between the means of the H_0, H_1 sampling distributions.

[2] The theory we present here is a specialization of Iain Embrey's 'States of Nature and States of Mind' formulation.[9]

mulation 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 attacker's decision over which type of phishing ploy to adopt takes on importance, as does their execution of that ploy, and their decision over how and when to attack. This observation is of strategic importance, because it implies that the attacker has first-mover advantage. Second, a single test-phishing email is insufficient for evaluating the effectiveness of email security training, because individuals' susceptibility 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. 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.

Our model's implications collectively suggest that an attacker could 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 do not involve special treatment of embedded links or attached files. 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 multiple dimensions over which an attacker could tailor a phishing campaign. For example, a sophisticated attacker can use rich contextual information to initiate a tailored *spear-phishing* campaign – i.e. to specifically target the 'routinely click-straight-through' choice criterion characterized by automaticity.

Taken together, these implications provide an explanation for the 'stepping-stone penetration pattern' that is common in APT attacks. In Section 4.3 we show that the stepping-stone pattern arises as a constrained-optimal attack vector under a model which admits both impulsive and automatic choice criteria. The lack of existing theoretical explanations for the stepping-stone attack vector may therefore be a consequence of the commonly maintained assumption that a single choice criterion can adequately describe all individuals on all occasions.

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, compares it to existing models (Section 4.2), and illustrates some of its properties. Section 4.3 demonstrates that the stepping-stone penetration pattern arises as a constrained-optimal attack vector, and Section 4.4 discusses our model's implications for security policy. Section 5 concludes.

2 Phishing Targets the Human Element

Human beings have the capacity for rational deliberation, but we do not always premeditate the full consequences of every action that we take. Indeed, social-engineering attacks are predicated upon the intuitive, emotional, and fallible nature of human behavior, and it is now recognized that psychology is an essential component of information security.[12]

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.[13] The FBI and the DHS recently issued an amber alert warning of APT activity targeting energy – especially, nuclear power [14] – and other sectors.[15] 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.[16]

Online scams such as phishing and spear phishinghave been labelled ‘social engineering attacks’ because they eschew direct, rational argumentation in favor of ‘peripheral’ routes to persuasion.[1, 5] 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.[1–6] Scams[3] typically augment peripheral-route persuasion by setting up a scenario that creates psychological pressure by triggering *visceral emotions* that override rational deliberation.[2, 17, 18] 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 both relief from the visceral discomfort as well as visceral satisfaction upon fulfilling the underlying need.

An ideally scripted scam scenario contrives a compelling, credible need for immediate action.[2, 17, 18] 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 distort the relative desirability of potential outcomes, but they also have a direct hedonic effect of narrowing and restricting attention toward a particular focal cue and its availability (or absence) in the present.[17, 18] Since visceral emotions and their effects are short lived, scam scripts emphasize a need for immediate action.

At sufficiently high levels of intensity, visceral emotions can override rational deliberation entirely.[17] 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

[3] as well as ‘hard-sell’ and ‘high-pressure’ marketing more generally,

necessarily stoking emotion. This approach targets the routinization and *automaticity*[11] 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 attached files – sent by bosses and immediate colleagues, and frequently also those sent by more distant contacts. Such unquestioning trust is expedient, until it is exploited by a spear phisher who can produce a plausible facsimile of an internal communication.

The spear-phishing strategy has proven effective at bypassing rational deliberation 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, Gmail, Lockheed Martin, Oak Ridge National Laboratory, RSA SecureID, and Wolf Creek Nuclear Operating Corporation.[13, 16, 19, 14]Employees at many of these institutions will be highly trained in information security, but a well-implemented spear-phishing attack with appropriate contextual information simply does not attract their critical evaluation.

3 Coexisting Choice Criteria: Empirical Provenance

A well-established body of empirical-decision-theory literature suggests that there may be more than one way to make any given 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, 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."[21] 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."[22] And using a Bayesian approach, Houser et al. show that different individuals use one of three different choice criteria when solving dynamic decision problems.[23]

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.[24] '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.[25] FM models are the technique of choice for analyzing Beauty-Contest data, revealing that virtually all 'non-game-theorist' subjects (94%) fall into one of three boundedly rational depth-of-reasoning classes (levels 0, 1 or 2).[26, 27] 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-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.,[28] 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, 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."[29] Preference heterogeneity documented in conventional single-choice-criterion models 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), 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).[29] 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.[30]

The results surveyed here mandate an interpretation of choice-criterion-selection probabilities that is nevertheless only implicit within the existing literature: (a) decision makers should not be characterized solely in terms of their *modal* choice criterion, but in terms of their choice-criterion mixtures, and (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. The explicit consideration of coexisting-choice-criteria is therefore a natural step toward improving the descriptive validity of theoretical models.

4 Incorporating *Homo intuivus, emotus, et fallibilis*

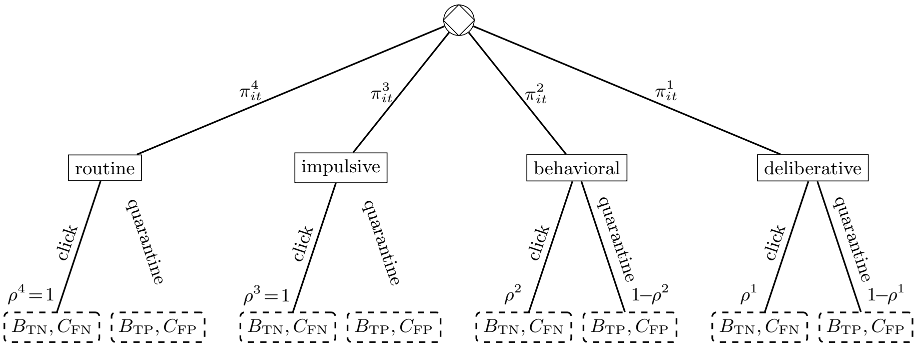
4.1 Coexisting-choice-criteria model

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

In the context of phishing-security, the elements of $\mathcal{C}$ must capture 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.[4]

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.


π^4_{it}
π^3_{it}
π^2_{it}
π^1_{it}
routine
impulsive
behavioral
deliberative
click
quarantine
click
quarantine
click
quarantine
click
quarantine
$\rho^4=1$
$\rho^3=1$
ρ^2
$1-\rho^2$
ρ^1
$1-\rho^1$
$B_{\mathrm{TN}}, C_{\mathrm{FN}}$
$B_{\mathrm{TP}}, C_{\mathrm{FP}}$
$B_{\mathrm{TN}}, C_{\mathrm{FN}}$
$B_{\mathrm{TP}}, C_{\mathrm{FP}}$
$B_{\mathrm{TN}}, C_{\mathrm{FN}}$
$B_{\mathrm{TP}}, C_{\mathrm{FP}}$
$B_{\mathrm{TN}}, C_{\mathrm{FN}}$
$B_{\mathrm{TP}}, C_{\mathrm{FP}}$

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).

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

[4] 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.[31] 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.[32] Automaticity, in which deliberative rationality is simply 'not engaged', has been given theoretical underpinning by Moors and De Houwer in the psychology literature,[11] and by P. Jehiel's concept of analogy-based expectations equilibrium in the economics literature.[36]

In general the choice-criterion selection probability will be conditional upon the decision maker's SoM, 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 $0 \leq \pi_{\mathrm{it}}^{c} \leq 1$ and $\sum_{c=1}^{C} \pi_{\mathrm{it}}^{c} = 1$ for all individuals i and time-points t.

Figure 1 illustrates a single agent's stochastic SoM 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 SoM 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.

In Section 4.3 we analyze an attacker's optimal response to the model illustrated in Figure 1, 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 $\boldsymbol{\alpha}_t$.

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

$$X_{it} = f\Big(\mathit{\Gamma}_i,\ \{Z_i\}_{<t}\,,\ \{\boldsymbol{\alpha}\}_{<t}\,,\ \{D_i\}_{<t}\Big)\ . \tag{1b}$$

The current characteristics X_{it} of agent i are jointly determined by their stable psychological traits $\mathit{\Gamma}_i$, and by the history of: decision contexts $\{Z_i\}_{<t}$, decision-attributes $\{\boldsymbol{\alpha}\}_{<t}$, and decision-outcomes $\{D_i\}_{<t}$, which jointly constitute 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[7] and we specialize the vectors appearing in (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 (see Section 4.4), 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 SoM. 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.[5] 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 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 agent i.

We may now extend the approach of Kaivanto[7] by defining the choice-criterion-specific match-quality function $m^c : [0,1]^A \times [0,1]^A \times \mathbb{R}^+ \to \mathbb{R}^+$ as

$$m_i^c(\boldsymbol{\alpha}) = m^c(\boldsymbol{\alpha}, \boldsymbol{s}_i^c, \chi_i^c) \qquad \forall\, c \in \mathcal{C} \ . \tag{2}$$

To illustrate, the simplest non-degenerate functional form for m^c is

$$m_i^c(\boldsymbol{\alpha}) = \chi_i^c + \boldsymbol{s}_i^c \cdot \boldsymbol{\alpha} \ , \tag{3}$$

where $\cdot$ denotes the vector dot product and m^c is linear-separable.

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

$$\pi_i^c(\boldsymbol{\alpha}) = \frac{m_i^c(\boldsymbol{\alpha})}{\sum_{c\in\mathcal{C}} m_i^c(\boldsymbol{\alpha})} \qquad \forall\, c \in \mathcal{C} \ . \tag{4}$$

4.2 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 $B_{\mathrm{TN}}, C_{\mathrm{FP}}, C_{\mathrm{FN}}$ and B_{TP}.[1, 3, 5, 6]

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 spear-phishing 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'

[5] The baseline propensity to adopt a deliberative choice criterion χ_i^1 is a stable trait[37] that can be measured by the CRT[35] or DMC scale[38].

names, responsibilities, 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).[1, 3–6] Importantly, there is variation between individuals in their susceptibility to particular levers of psychological influence.[6] For instance scarcity and authority have been found to be more effective for young users, while reciprocation and liking/affinity have been found to be more effective for older users.[6] 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.

It is important to note that none of the aforementioned psychological levers would be effective if email users were solely normatively rational deliberators. This observation suggests several empirically testable implications of the coexisting-choice-criteria model of security behavior.

First, we might seek to verify that individuals are more able to detect phishing emails when they are specifically asked to do so within a known simulated test, than when they are blind to being asked to do so. Such a result would contradict the prediction of any theory predicated upon a single choice criterion, since the consequences of failing to detect a phishing email are far more substantial outside of a known simulated test than they are within it.

Second, we might seek corroborative evidence for the hypotheses that π_{it} is individual- and situation-specific. This could be achieved by designing a battery of test phishing messages $\{\alpha^3\} \cup \{\alpha^4\}$ that separately target impulsive and routine choice criteria, and by measuring individuals' susceptibility to each of these within both blind and non-blind test designs. The hypotheses would be maintained if the above difference in susceptibility between blind and non-blind tests varies systematically (i) between individuals, and (ii) within individuals between $\{\alpha^3\}$ and $\{\alpha^4\}$.

Third, we might seek to move beyond such corroborative evidence to establish causality. This could be achieved by randomly assigning individuals to alternative training treatments, each of which targets a specific and distinct choice-criterion susceptibility. If experimental variation in blind cf. non-blind test susceptibility is induced, then we would have robust evidence in support of the theory proposed here. We suggest that future work that implements some or all of these tests would make an important contribution to the literature.

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 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

$$\max_{\boldsymbol{\alpha}_k} \quad \sum_{i=1}^{m}\sum_{c=1}^{C} \pi_i^c(\boldsymbol{\alpha}_k)\cdot\rho_i^c\cdot V \;\; - \; e(\boldsymbol{\alpha}_k) \qquad \text{s.t.} \qquad ||\boldsymbol{\alpha}_k|| \leq 1 \;\; , \tag{5}$$

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.[39, 40]

The attacker may send one or more, emails $\boldsymbol{\alpha}_k$. Each email may be designed to induce one particular SoM 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 characterize the available pure strategies on the basis of the phishing literature,[1, 3, 5] 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 (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 strictly dominated by the Impulsive strategy whenever

$$\rho^2 < \frac{\max_{\boldsymbol{\alpha}}\{\pi^3\}}{\max_{\boldsymbol{\alpha}}\{\pi^2\}} \ , \tag{6}$$

i.e. whenever the expected click-through probability under a Behavioral choice criterion is less than the relative ease of invoking the Behavioral SoM compared to the Impulsive SoM. 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:

$$\mathrm{P}\left({\text{non-target} \atop \text{clickthrough}}\right) = 1-(1-\max_{\boldsymbol{\alpha}}\{\pi^3\})^{n-m} \quad > \quad 1-(1-\max_{\boldsymbol{\alpha}}\{\pi^3\})^{m} = \mathrm{P}\left({\text{target} \atop \text{clickthrough}}\right)$$

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

$$\frac{e\big(\mathrm{argmax}_{\boldsymbol{\alpha}}\{\pi^4\}\big)}{e\big(\mathrm{argmax}_{\boldsymbol{\alpha}}\{\pi^3\}\big)} < \frac{\max_{\boldsymbol{\alpha}}\{\pi^4\}}{\max_{\boldsymbol{\alpha}}\{\pi^3\}} \ . \tag{7}$$

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 (6), (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.

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.[41] 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 ISOs should actively engage with wider aspects of organizational culture.

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 ISOs is that the vulnerabilities ρ^3 and ρ^4 are essentially fixed at 1. Once an Impulsive or Routine SoM 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,[35] 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 an Impulsive or Routine SoM.

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, testing 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,[7, 8] 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 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 (ISOs). 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 SoM c and the probability of then clicking through *given* this SoM. Hence training must address the different SoM 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 SoM choice criteria. In light of the coexisting-choice-criteria model, the single-test-email approach should be deprecated.

The coexisting-choice-criteria model highlights 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 ISOs could – and perhaps should – be afforded a broader role in shaping organizational culture.

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

References

1. Rusch, J.J.: The "social engineering" of internet fraud. Proceedings of the Internet Society Global Summit (INET'99), 1999, June 22–25, San Jose, CA.
2. Langenderfer, J., Shimp, T.A.: Consumer vulnerability to scams, swindles, and fraud: A new theory of visceral influences on persuasion. Psyc. Mark. **18**,763–783 (2001).
3. Mitnick, K.D., Simon, W.L.: The Art of Deception: Controlling the Human Element of Security. Indianapolis, IN: Wiley, 2002.
4. Cialdini, R.B.: Influence: The Psychology of Persuasion. New York, NY: Collins, 2007.
5. Hadnagy, C.: Social Engineering: The Art of Human Hacking. Indianapolis: Wiley, 2011.
6. Oliveira, D., et al.: Dissecting spear phishing emails for older vs young adults. Proc. 2017 CHI Conf. Hum. Factors Comp. Sys. 6412–6424 (2017).
7. Kaivanto, K.: The effect of decentralized behavioral decision making on system-level risk. Risk Anal. **34**(12), 2121–2142 (2014).
8. Canfield, C.I., Fischhoff, B.: Setting priorities for behavioral interventions: An application to reducing phishing risk. Risk Anal. **38**(4), 826–838 (2018).
9. Embrey I.: States of nature and states of mind: A generalised theory of decision-making. Working Paper 2017/032, Lancaster Univ. Econ. Dept. (2017).
10. Tversky, A., Kahneman, D.: Advances in prospect theory: Cumulative representation of uncertainty. J. Risk Unc. **5**(4), 297–323 (1992).
11. Moors A, DeHouwer J.: Automaticity: A theoretical and conceptual analysis. Psych. Bull. **132**(2), 297–326 (2006).
12. Anderson, R., Moore, T.: Information security: Where computer science, economics and psychology meet. Phil. Trans. Roy. Soc. A **367**(1898), 2717–2727 (2009).
13. Johnson, N.: Feds' chief cyberthreat: 'Spear phishing' attacks. FedT, 2/20, (2013).
14. Perlroth, N.: Hackers are targeting nuclear facilities, Homeland Security Dept. and F.B.I. say. New York Times, July 6, 2017.
15. FBI and DHS. Advanced Persistent Threat Activity Targeting Energy and Other Critical Infrastructure Sectors. 'Amber' Alert (TA17-293A), Oct 20, 2017.
16. Elgin, B., Lawrence, D., Riley, M.: Coke gets hacked and doesn't tell anyone. Bloomberg, Nov 4, 2012.
17. Loewenstein, G.: Out of control: Visceral influences on economic behavior. Org. Beh. Hum. Perf. **65**(3), 272–292 (1996).
18. Loewenstein, G.: Emotions in economic theory and economic behavior. Am. Ec. Rev. **90**(2), 426–432 (2000).
19. Hong, J.: The state of phishing attacks. Comm. ACM **55**(1), 74–81 (2012).
20. Kahneman, D.: Thinking, Fast and Slow. New York, NY: Penguin, 2012.
21. Harrison, G.W., Rutström, E.E.: Expected utility theory and prospect theory: One wedding and a decent funeral. Exp. Econ. **12**(2), 133–158 (2009).
22. Coller, M., Harrison, G.W., Rutström, E.E.: Latent process heterogeneity in discounting behavior. Oxf. Ec. Pap. **64**(2), 375–391 (2011).
23. Houser, D., Keane, M., McCabe, K.: Behavior in a dynamic decision problem. Econometrica **72**(3), 781–822 (2004).
24. Stahl, D.O., Wilson, P.W.: On players' models of other players: Theory and experimental evidence. Games Ec. Beh. **10**(1), 218–254 (1995).

25. Nagel, R.: Unraveling guessing games: An experimental study. Am. Ec. Rev. **85**(5), 1313–1326 (1995).
26. Stahl, D.O.: Boundedly rational rule learning in a guessing game. Games Ec. Beh. **16**(2), 303–313 (1995).
27. Bosch-Domènech, A., Montalvo, J., Nagel, R., Satorra, A.: A finite mixture analysis of beauty-contest data using generalized beta distributions. Exp Ec **13**, 461-475 (2010).
28. Rothschild, C., McLay, L., Guikema, S.: Adversarial risk analysis with incomplete information: A level-*k* approach. Risk Anal. **32**(7), 1219–1231 (2012).
29. Hess, S., Stathopoulos, A., Daly, A.: Allowing for heterogeneous decision rules in discrete choice models: An approach and four case studies. Transp. **39**, 565-591 (2012).
30. Swait, J., Adamowicz, W.: The influence of task complexity on consumer choice: A latent class model of decision strategy switching. J. Cons. Res. **28**, 135–148 (2001).
31. Wakker, P., Tversky, A.: An axiomatization of cumulative prospect theory. J. Risk Unc. **7**(2), 147–175 (1993).
32. Frank, R.H.: Passions Within Reason: The Strategic Role of the Emotions. New York, NY: Norton, 1988.
33. Chou, E., McConnell, M., Nagel, R., Plott, C.R.: The control of game form recognition in experiments. Exp. Ec. **12**(2), 159–179 (2009).
34. VanLehn, K.: Mind bugs: The Origins of Procedural Misconceptions. Cambridge, MA: MIT, 1990.
35. Frederick, S. Cognitive reflection and decision making. J. Ec. Persp. **19**, 25-42 (2005).
36. Jehiel, P.: Analogy-based expectation equilibrium. J. Ec. Th. **123**, 81–104 (2005).
37. Parker, A.M., De Bruin, W.B., Fischhoff, B., Weller, J.: Robustness of decision-making competence. J. Beh. Dec. Mak. **31**(3), 380–391 (2018).
38. Parker, A.M., Fischhoff, B.: Decision-making competence: External balidation through an individual-differences approach. J. Beh. Dec. Mak. **18**(1), 1–27 (2005).
39. Shapiro, C, Varian, H.R.: Information Rules: A Strategic Guide to the Network Economy. Boston, MA: Harvard Business School Press, 1998.
40. Anderson, R.J.: Security Engineering: A Guide to Dependable Distributed Systems. Indianapolis, IN: John Wiley, 2008.
41. Steigenberger, N., Lübcke, T., Fiala, H., Riebschläger, A.: Decision Modes in Complex Task Environments. London: CRC Press (Taylor & Francis), 2017.