The Effect of Decentralized Behavioral Decision Making on System-Level Risk*

Kim Kaivanto†

Department of Economics, Lancaster University, Lancaster LA1 4YX, UK

this version: February 20, 2014

†tel +44(0)1524594030; fax +44(0)1524594244; e-mail k.kaivanto@lancaster.ac.uk

Abstract

Certain classes of system-level risk depend partly on decentralized lay decision making. For instance, an organization's network security risk depends partly on its employees' responses to phishing attacks. On a larger scale, the risk within a financial system depends partly on households' responses to mortgage sales pitches. Behavioral economics shows that lay decision makers typically depart in systematic ways from the normative rationality of Expected Utility (EU), and instead display heuristics and biases as captured in the more descriptively accurate Prospect Theory (PT). In turn psychological studies show that successful deception ploys eschew direct logical argumentation and instead employ peripheral-route persuasion, manipulation of visceral emotions, urgency, and familiar contextual cues. The detection of phishing and inappropriate mortgages may be framed as a binary classification task. Signal Detection Theory (SDT) offers the standard normative solution, formulated as an optimal cutoff threshold, for distinguishing between good/bad emails or mortgages. In this paper we extend SDT behaviorally by re-deriving the optimal cutoff threshold under PT. Furthermore we incorporate the psychology of deception into determination of SDT's discriminability parameter. With the neo-additive probability weighting function, the optimal cutoff threshold under PT is rendered unique under well-behaved sampling distributions, tractable in computation, and transparent in interpretation. The PT-based cutoff threshold is (i) independent of loss aversion and (ii) more conservative than the classical SDT cutoff threshold. Independently of any possible misalignment between individual-level and system-level misclassification costs, decentralized behavioral decision makers are biased toward *under-detection*, and system-level risk is consequently greater than in analyses predicated upon normative rationality.

Keywords: system-level risk; Signal Detection Theory; Prospect Theory; psychology of deception; spear phishing

JEL classification: D81

1 INTRODUCTION

Computer networks and securitization markets are examples of systems in which the self-interested actions of lay decision makers contribute to the severity of system-level risk. An organization's network security may be compromised from the staging ground of individual user accounts. Hence overall network security risk depends on individual users' decision making in the face of phishing attacks.[1] And in markets for securitized mortgage products, the potential loss distribution depends in part on individual homebuyers' decisions to accept or reject high-risk mortgage contracts.

So although much can be done to limit system-level risk through alignment of incentives and technical and procedural protocols, ultimately it is a collection of lay people – employees or customers – whose individual decisions carry system-level consequences. To render these decisions amenable to modeling and incorporation into a formal risk model, it is useful to frame the individual-level decisions as binary classification tasks – between authentic and malicious emails, or between appropriate and inappropriate mortgage contracts.

In the Signal Detection Theory (SDT) formalization of binary classification, the signal extracted by an individual is represented by the magnitude of a score variable. Higher values of this score variable are associated with the malicious/inappropriate class. Where the sampling distribution of a score variable is known both under the null (benign) hypothesis as well as under the alternative (malicious) hypothesis, classical SDT identifies the optimal cutoff threshold in this score variable for binary classification by minimizing the expected cost of misclassification errors, striking the optimal trade-off between the true positive likelihood and the false positive likelihood from among the set of feasible combinations. The frontier of the latter feasible set is known as a Receiver Operating Characteristics (ROC) curve.

This analytical machinery is useful, particularly as a benchmark for individuals who conform with normative decision theory. However, ample and robust experimental evidence shows that most individuals do not conform with normative decision theory, and instead display a variety of heuristics and biases.(1) In this paper we re-derive the SDT optimal cutoff threshold under the more descriptively accurate objective function of Prospect Theory (PT).(2)

The resulting new form of the optimal cutoff threshold identifying expression differs from

[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.

its classical SDT precursor. Loss aversion does not appear in this identifying expression. But due to the PT value function, the ratio of the subjective impact of the cost of Type I error to the subjective impact of the cost of Type II error is greater in the PT-based model than in the classical model. This leads to a more conservative cutoff threshold having smaller true positive and false positive likelihoods.

But the most striking features are due to the non-linear probability weighting function. This (i) introduces non-linearity into the objective function's contours, resulting in potential non-uniqueness of optimal cutoff threshold recommendations, and (ii) renders the cutoff threshold expression opaque to interpretation and intractable for non-computer-intensive calculation. Both of these complications are resolved by the use of a linear-with-boundary-discontinuities 'neo-additive' probability weighting function.

Compared with PT-based behavioral decision makers, assuming instead that individual network users abide by normative decision theory entails *underestimation* of system-level risk. When the model is extended to account for the psychology of deception, this underestimation of system-level risk is even more pronounced. To illustrate the magnitude and consequentiality of this underestimation for system-level risk, we develop an SDT-based phishing-risk model, which we evaluate using agent-based simulation modeling under three sets of assumptions: non-behavioral SDT, PT-SDT, and PT-SDT incorporating psychology-of-deception effects. These simulation results reinforce and extend the insights from comparative static analysis, and furthermore showcase the potential for calibrated variants of the type of modeling apparatus developed here to be used instrumentally by Information Security Officers for security-breach risk estimation.

Henceforth in this paper, the development is couched in terms of the email classification problem. Full treatment of the mortgage classification problem, employing the method and results developed here, is deferred to subsequent work.

2 PHISHING LITERATURE PRÉCIS

This work bridges the interstices between several strands of literature, some of which are in their infancy. The model itself is developed within the mature SDT framework (see Section 3). This is 'behavioralized' by substituting the classical optimal cutoff threshold with one derived under Tversky and Kahneman's PT objective function (see Section 4). PT, which formally is a

generalization of Expected Utility (EU), incorporates a number of the principal findings of the 'heuristics and biases' literature: framing, nonlinear probability weighting, source dependence, risk aversion in gains, risk seeking in losses, loss aversion, ambiguity aversion, and the four-fold pattern of risk attitudes.[1,2] Within the computer science literature, Ryan West is credited with an early exploration of the implications of PT for security,[3] closely followed by Ross Anderson and Tyler Moore's influential characterization of information security as a field at the intersection between computer science, economics and psychology.[4]

West employs PT as a lens with which to identify factors that have a bearing on the trade-offs between security risks, losses and benefits.[3] Pro-security actions typically involve sure immediate costs (e.g. inconvenience, delay). Meanwhile, the costs of security breaches are uncertain and occur in the future. West observes that, due to the properties of the PT value function, behavioral decision makers are more likely to gamble on the possibility of not having to face a security breach rather than incur the sure immediate cost of a pro-security action. And given that PT also embodies 'loss aversion' – i.e. that losses loom larger than gains, generally by a factor in excess of 2 – system designers must recognize that although the inconvenience cost of a security measure may be small, loss averse users will require an off-setting benefit more than twice as large in order to render the security measure psychologically worthwhile. Thus, when examined through the lens of PT, Cormac Herley's finding – that "most security advice simply offers a poor cost-benefit tradeoff to users and is rejected"[5] – does not go far enough. For loss averse behavioral users, even costs balanced one-for-one with benefits are unattractive, and remain so until the cost-benefit ratio drops below $\frac{1}{2}$.

For obvious reasons, much of the research on phishing is computer-science centric.[6] But the vulnerability that phishing and other 'social engineering' hacks exploit is the human user, who is a psychological and behavioral agent, rather than axiomatically rational. Thus the thrust of research has increasingly turned toward online trust,[7] detection of deception,[8–10] and phenomena such as dynamic inconsistency induced by hyperbolic discounting.[11]

Online scams such as phishing employ the social engineering techniques of persuasion and influence. Rather than the direct, rational argumentation route to persuasion, scams follow a peripheral route to persuasion that largely bypasses logical processes. Research in psychology has identified at least six different factors that may be deployed in peripheral-route persuasion: authority, scarcity, similarity and identification, reciprocation, consistency following commit-

ment, and social proof.[12–14] As in legitimate forms of marketing, scams emphasize the urgency of the opportunity or required action.[15] If urgency is taken at face value, then there is not time to contemplate and 'think on it' as one would in the direct route to persuasion. Furthermore, scammers invoke a subset of *visceral factors* to override rational deliberation and increase the relative desirability of compliance.[15,16] Emotions such as greed, pity, lust, fear and anxiety are visceral factors which, once stirred up, act as a 'stick' if the associated need is not met, and provide a 'carrot' when the need is met. For instance, emails purportedly from the IRS which inform the recipients of 'Unreported/Underreported Income (Fraud Application)' are designed to trigger fear and anxiety. But this emotionally charged, viscerally motivated state does not persist long. Hence scams invariably contrive compelling reasons for immediate action.[16] In the words of a former swindler "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."[17]

Embedding the phishing ploy within an email containing rich recipient-specific contextual information has become possible as digital footprints have grown. This targeted and tailored approach is called *spear phishing*, and it is typically the first stage of an Advanced Persistent Threat (APT) attack on an organization's sensitive information. Numerous organizations – governmental, defense, corporate and scientific – have been compromised in this manner, including 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., and Chesapeake Energy.[18–21] The contextual information in spear-phishing emails enhances the effectiveness of appeals to authority, credibility, similarity & identification, reciprocation, consistency, and social proof. Evidence for this enhancement comes not only from the success of spear phishing 'in the field', but also from controlled experiments. Jagatic et al.[22] for instance found that merely spoofing emails so as to appear to be sent by an individual that the recipient recognizes causes a 4.5-fold increase in the susceptibility to clicking a link in the malicious email. When implemented well, a spear-phishing email seemingly does not stand out among the target's legitimate emails.

What this literature précis reveals is that behavioral decision makers differ from their normative counterparts on two levels: (i) their trade-offs between uncertainties, losses and benefits are descriptively captured by PT rather than EU, and (ii) their susceptibility to peripheral-route persuasion, visceral emotions, and disarming by familiar cues introduces gaps into information

processing, possibly even shifting the locus of attention away from the key question of benign/malicious content entirely. The former (i) is addressed below by re-deriving SDT's optimal cutoff threshold under the PT objective function (see Section 4). The latter (ii) is addressed by re-expressing the ROC curve as a function of peripheral-route persuasion (see Section 5).

3 CLASSICAL SIGNAL DETECTION THEORY

A binary classifier is sought for whether malicious content is present (D) or absent ($\neg D$). Following the standard SDT formulation, the problem reduces to the determination of an optimal cutoff threshold $\theta^* \in \Theta = [\underline{\theta}, \overline{\theta}] \subset \mathbb{R}$ that identifies the observed score θ as belonging either to the acceptance interval ($\underline{\theta} \leq \theta \leq \theta^*$) associated with acceptance of the null hypothesis $H_0 : \neg D$ or to the rejection interval ($\theta^* < \theta \leq \overline{\theta}$) in which the null hypothesis is rejected in favor of the the alternative hypothesis $H_1 : D$. The optimal cutoff threshold is identified by applying an optimality criterion – e.g. minimizing expected cost, or maximizing expected utility – subject to the error likelihoods being constrained by the ROC curve.[23–25] Given that the scoring procedure yields a different sampling distribution for θ under the null than under the alternative, different cutoff thresholds θ' yield different Type I and Type II error likelihoods (α, β).

Table 1: Likelihood assignments and associated terminology

$\text{TNL}_{\theta'}$	$= P(\theta \leq \theta' \mid \neg D)$	$= (1 - \alpha_{\theta'})$	= 'Specificity'
$\text{FPL}_{\theta'}$	$= P(\theta > \theta' \mid \neg D)$	$= \alpha_{\theta'}$	= Type I error likelihood
$\text{TPL}_{\theta'}$	$= P(\theta > \theta' \mid D)$	$= (1 - \beta_{\theta'})$	= 'Sensitivity'; Power
$\text{FNL}_{\theta'}$	$= P(\theta \leq \theta' \mid D)$	$= \beta_{\theta'}$	= Type II error likelihood

For every scoring procedure each particular threshold value θ' defines a combination of True Negative Likelihood (TNL), False Positive Likelihood (FPL), True Positive Likelihood (TPL) and False Negative Likelihood (FNL), where the former pair and the latter pair are complementary ($\text{FPL}_{\theta'} = 1- \text{TNL}_{\theta'}$ and $\text{TPL}_{\theta'} = 1- \text{FNL}_{\theta'}$). Given that the null and alternative hypotheses are operationalized as $H_0 : \theta \leq \theta'$ and $H_1 : \theta > \theta'$, the correspondences in Table 1 hold.

The ROC curve for a scoring procedure plots the $\text{TPL}_{\theta'}$ on the vertical axis of the unit square against the $\text{FPL}_{\theta'}$ on the horizontal axis of the unit square as the threshold θ' is varied within its domain (see Figure 1a). In other words the ROC curve consists of the parametric plot of

($\text{FPL}_{\theta'}$,$\text{TPL}_{\theta'}$) which results when the cutoff threshold is allowed to vary within the support of the score variable $\{(P(\theta > \theta'|\neg D), P(\theta > \theta'|D)) : \theta' \in \Theta\}$. Sampling distributions that coincide everywhere $f(\theta|\neg D) = f(\theta|D)\ \forall\, \theta \in \Theta$ yield a classifier that performs no better than chance; the ROC curve of this classifier coincides with the diagonal. When the sampling distributions of θ under $\neg D$ and D are unimodal, continuously differentiable and $F(\theta|\neg D) \leq F(\theta|D)\ \forall\, \theta \in \Theta$, the ROC curve is everywhere differentiable and monotonically decreasing in slope. The Area Under the Curve (AUC) ranges from $\frac{1}{2}$ for the random classifier to AUC=1 for a perfectly discriminating classifier.

Where the score variable is normally distributed with common variance in both negative $\theta \sim N(\mu_{\neg D}, \sigma^2)$ and positive $\theta \sim N(\mu_D, \sigma^2)$ states with $\mu_D \geq \mu_{\neg D}$, then the *discriminability index* is defined as

$$d' = \frac{\mu_D - \mu_{\neg D}}{\sigma} \quad . \tag{3.1}$$

The greater the distance between the means of the sampling distributions, the more discriminating the signal and the larger the AUC. As $d' \to 0$, AUC $\to \frac{1}{2}$; and AUC $\to 1$ as $d' \to \infty$.

Denoting the direct cost of implementing the generic scoring procedure as C and the costs associated with true positives, false negatives, true negatives and false positives as C_{TP}, C_{FN}, C_{TN} and C_{FP} respectively, then the expected cost of using the signal detection mechanism is of the form $E(C) = C + C_{\text{TP}}P(\text{TP}) + C_{\text{FN}}P(\text{FN}) + C_{\text{TN}}P(\text{TN}) + C_{\text{FP}}P(\text{FP})$. ROC curves are continuous, but need not be everywhere differentiable. For ROC curves that are differentiable, θ^* identifies the point (FPL_{θ^*}, TPL_{θ^*}) at which the iso-expected-cost line is tangent to the ROC curve. We wish to minimize the expected costs of implementing the decision criterion $E(C)$ subject to the TPL and FPL parameters being constrained by the ROC curve, which for present purposes we represent as the function $\text{TPL} = G(\text{FPL})$.

$$\min_{\theta'}\ E(C) \qquad \text{s.t.} \qquad \text{TPL} = G(\text{FPL}) \tag{3.2}$$

The slope of each iso-expected-cost contour – and therefore also the slope of the cost minimizing iso-expected-cost line at the optimal operating point – is the ratio of expected opportunity cost of misclassifying an authentic email ($\neg D$) to the expected opportunity cost of misclassifying a

malicious email (D).

$$\left(\frac{\text{dTPL}}{\text{dFPL}}\right)_{\bar{C}} = \frac{P(\neg D)}{P(D)}\left[\frac{C_{\text{FP}} - C_{\text{TN}}}{C_{\text{FN}} - C_{\text{TP}}}\right] = \left(\frac{\text{dTPL}}{\text{dFPL}}\right)_{\bar{C}^*} \quad . \tag{3.3}$$

The optimal cutoff threshold is the θ^* that generates the point on the ROC curve $(\text{FPL}_{\theta^*}, \text{TPL}_{\theta^*}) = (P(\theta > \theta^*|\neg D), P(\theta > \theta^*|D))$ which satisfies the tangency condition (3.3).

The square-bracketed term in (3.3) fixes the manner in which misclassification costs affect the optimal cutoff threshold. Only the cost difference between the misclassification and the correct classification matters for optimal cutoff threshold placement. Not, for instance, cost differences across the authentic/malicious state divide. Similarly, the *levels* of within-state costs are cutoff-threshold irrelevant; only their difference matters, via the ratio across states.

The slope of the iso-expected-cost contour (3.3) is also the optimal critical likelihood ratio l^* with which to assign observed scores θ either to the acceptance region or the rejection region according to the likelihood ratio condition: if $l_{D,\neg D}(\theta) = \frac{P(\theta|D)}{P(\theta|\neg D)} < l^*$ then $H_0 : \neg D$ is accepted, or if $l_{D,\neg D}(\theta) = \frac{P(\theta|D)}{P(\theta|\neg D)} \geq l^*$ then $H_0 : \neg D$ is rejected in favour of $H_1 : D$.[24]

Classical signal detection theory recognizes that, in general, the cost terms can be replaced by the utilities of incurring such costs. Nevertheless this possibility is neither widely explored nor widely adopted within the broader literature.[2] The standard classical approach is to minimize expected cost.

4 APPLICATION OF PT TO SDT

4.1 Setup

Consider the generic prospect $(\boldsymbol{x}, \boldsymbol{p})$, composed of $m+n+1$ outcomes $\boldsymbol{x}_{(m+n+1\times 1)} = (x_{-m}, ..., x_0, ..., x_n)$ where $x_{-m} < \cdots < x_0 < \cdots < x_n$ and probabilities $\boldsymbol{p}_{(m+n+1\times 1)} = (p_{-m}, ..., p_0, ..., p_n)$. Under PT the preference value of a prospect $(\boldsymbol{x}, \boldsymbol{p})$ is given by $V(\boldsymbol{x}, \boldsymbol{p}) = V^+(\boldsymbol{x}, \boldsymbol{p}) + V^-(\boldsymbol{x}, \boldsymbol{p})$ where V^+ and V^- are the contributions of gains and losses respectively.

In the present framework we will be concerned exclusively with the (mis)classification costs

$$C_{\text{FN}} > C_{\text{TP}} > C_{\text{FP}} > C_{\text{TN}} \geq 0 \quad , \tag{4.1}$$

[2] Two exceptions may be noted. Ulehla[26] and Galanter[27] "have proposed that deviations from the normative prescriptions arise for asymmetrical payoff matrices when the utility of money is a negatively accelerated function of the monetary values included in the matrix".[28]

where 'Negative' denotes the benign classification ($\neg D$) and 'Positive' denotes the malicious classification (D). It is natural to set $C_{\mathrm{TN}} = 0$, as there are no follow-on 'costs' to opening a non-malicious email. It is also natural that C_{FN} is the largest element in the set of misclassification costs, as failing to detect and reject a malicious email has the worst possible consequences in this context. In-between, the ranking of C_{TP} and C_{FP} is unambiguous in targeted and tailored 'spear phishing' forms of attack.[18–22] Here False Positives involve non-zero costs; beyond secure deletion, protocol requires they be reported to and investigated by network security personnel. Similarly True Positives must also be reported and investigated, but in this case the procedure is more intrusive, disruptive and protracted.

This cost structure is shared by most signal detection problems where identification of the D state is consequential to the organism, individual, or organization. This is the case for instance in medical diagnosis. Furthermore, as Michael Shermer has argued with regard to animal and human evolution more broadly, false positives – believing that there is a connection between A and B when there is not – are usually harmless; in contrast, false negatives – believing that there is no connection between A and B when there is – may have life- and procreation-threatening consequences.[29] Note that, as this paper's objective is to study the effects of the behavioral nature of users, we abstract from any possible misalignment of incentives (i.e. misalignment of misclassification costs) between users and the organization as a whole.

Due to (4.1), the present analysis is conducted entirely within the loss domain, where the PT preference value of a prospect is given by

$$V^{-}(\boldsymbol{x},\boldsymbol{p}) = w^{-}(p_{-m})v^{-}(x_{-m}) + \sum_{k=1}^{m}\left[w^{-}\left(\sum_{j=0}^{k} p_{-(m-j)}\right) - w^{-}\left(\sum_{j=0}^{k-1} p_{-(m-j)}\right)\right] v^{-}(x_{-(m-k)}) \quad (4.2)$$

In accordance with widespread practice within the PT literature,[30–32] we follow Tversky and Kahneman[2] (TK92) in adopting a power-function specification of the value function

$$v^{-}(x) = -\lambda\cdot(-x)^{\phi^{-}} \quad \text{for} \quad x \leq 0, \ \text{ with } \ \phi^{-} = 0.88 \ \text{ and } \ \lambda = 2.25 \ , \quad (4.3)$$

and the single-parameter probability weighting function

$$w^{-}(p) = \frac{p^{\delta}}{(p^{\delta} + (1-p^{\delta}))^{1/\delta}} \qquad \text{with } \ \delta = 0.69 \ . \quad (4.4)$$

This combination of functional forms and parameters has been obtained through maximum likelihood estimation applied to laboratory experiment data. However, these TK92 functional forms and parameters have also received support from 'parameter-free' elicitation procedures.[39] They are the initial defaults used in applications of PT.[3]

4.2 Optimal operating point under PT

As above, the email is either malicious (inappropriate) with prior probability $p = P(D)$ or is authentic with prior probability $1 - p = P(\neg D) = 1 - P(D)$. Conditional on the true state, the scoring mechanism generates, for any given threshold $\theta' \in [\underline{\theta}, \overline{\theta}]$, the classifications TN, FP, TP, FN according to the probabilities outlined in Table 1 above. Then the PT preference function (over the loss domain) simplifies to

$$\begin{aligned} V^-(C) \;\; = \;\; & - \; w^-(p\beta)\lambda[v^-(C_{\mathrm{FN}}) - v^-(C_{\mathrm{TP}})] \; - \; w^-(p)\lambda[v^-(C_{\mathrm{TP}}) - v^-(C_{\mathrm{FP}})] \\ & - \; w^-(p + (1-p)\alpha)\lambda[v^-(C_{\mathrm{FP}}) - v^-(C_{\mathrm{TN}})] \; - \; w^-(1)\lambda v^-(C_{\mathrm{TN}}) \;\; . \end{aligned} \tag{4.5}$$

Solving for the slope of iso-$V^-(C)$ contours in ROC space (see Appendix A for the $\left(\frac{\psi_1(\alpha,\beta|p,\delta)}{\psi_2(\alpha,\beta|p,\delta)}\right)$ term):

$$\frac{\mathrm{dTPL}}{\mathrm{dFPL}} = \left[\frac{(C_{\mathrm{FP}})^{\phi^-} - (C_{\mathrm{TN}})^{\phi^-}}{(C_{\mathrm{FN}})^{\phi^-} - (C_{\mathrm{TP}})^{\phi^-}}\right] \cdot \left(\frac{\psi_1(\alpha,\beta|p,\delta)}{\psi_2(\alpha,\beta|p,\delta)}\right) \tag{4.6}$$

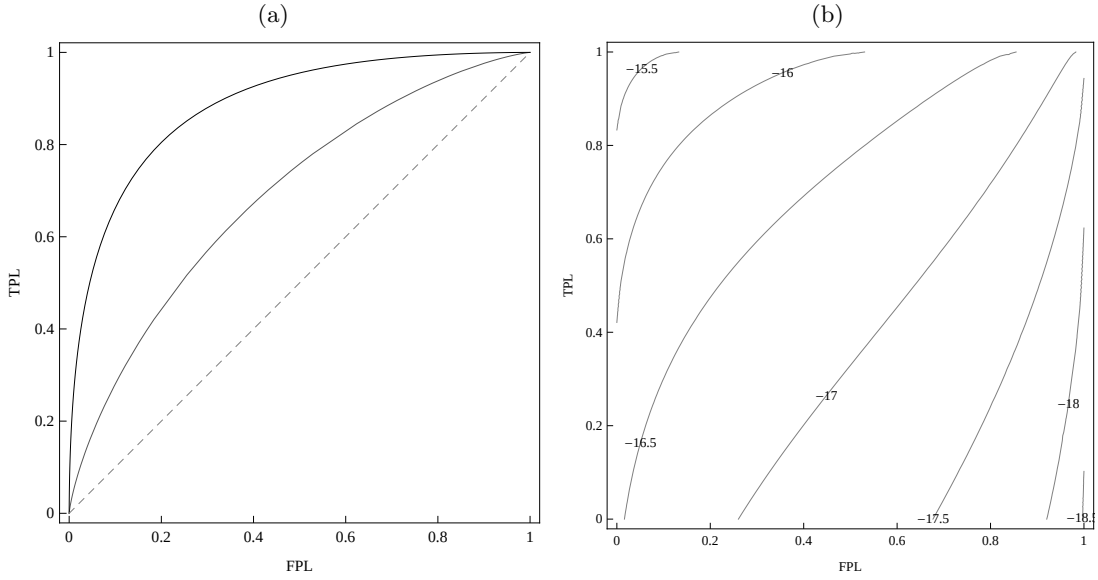
There are two terms and they enter multiplicatively. Firstly, in the square brackets, the ratio of the penalty incurred for misclassification of authentic emails to the penalty incurred for misclassification of malicious emails. Here the 'penalty' is gauged by the value function over negative outcomes (i.e. costs) calculated with regard to the reference point $C_{\mathrm{TN}} = 0$. Secondly, in the round brackets, a ratio term in probabilities, which here reflects not only prior probability p, but also *weighting* of the prior probability and misclassification likelihoods. The form of this (weighted) probability term is complex; it resists simplification, intuitive interpretation, as well as manual computation. Hence in (4.6) the term $\left(\frac{\psi_1(\alpha,\beta|p,\delta)}{\psi_2(\alpha,\beta|p,\delta)}\right)$ stands as a place holder for the complete expression presented in Appendix A.

Using numerical methods it may be shown that the iso-$V^-(C)$ contours vary over a range from nearly horizontal to nearly vertical, depending on the parameter values. At the near-vertical

[3] `http://prospect-theory.behaviouralfinance.net/cpt-calculator.php`, `http://psych.fullerton.edu/mbirnbaum/calculators/cpt_calculator.htm`

and near-horizontal extremes, the contours are have less curvature and are approximately linear. In the intermediate range however, the contours are characterized by pronounced curvature, as illustrated in Figure 1b. This curvature of the iso-$V^-(C)$ contours introduces the possibility, depending on the shape of the ROC curve, of non-uniqueness of the (FPL,TPL)= $(\alpha, (1-\beta))$ point identified as optimal.

Figure 1: (a) ROC curves generated from $\theta \sim N(1,1)$ under H_0 and $\theta \sim N(1.7,1)$ and $\theta \sim N(2.7,1)$ under H_1. (b) Typical intermediate-range iso-$V^-(C)$ contours under the TK92 probability weighting function.



Note that the loss aversion parameter λ cancels out of the iso-$V^-(C)$ contour slope expression (4.6).

The square bracketed classical misclassification cost ratio term in equation (3.3) is *smaller* than the corresponding square-bracketed term in (4.6). Recall that $C_{\mathrm{TN}} = 0$, and define the remaining cost terms with the constants $\{c_i \in (1,\infty] \subset \mathbb{R},\ i = 2,3,4\}$ such that $C_{\mathrm{TN}} = 0 < C_{\mathrm{FP}} = c_2 < C_{\mathrm{TP}} = c_2c_3 < C_{\mathrm{FN}} = c_2c_3c_4$. Recalling that $\phi^- < 1$, it follows that

$$\left[\frac{(C_{\mathrm{FP}})^{\phi^-} - (C_{\mathrm{TN}})^{\phi^-}}{(C_{\mathrm{FN}})^{\phi^-} - (C_{\mathrm{TP}})^{\phi^-}}\right] \quad > \quad \left[\frac{C_{\mathrm{FP}} - C_{\mathrm{TN}}}{C_{\mathrm{FN}} - C_{\mathrm{TP}}}\right] \quad . \tag{4.7}$$

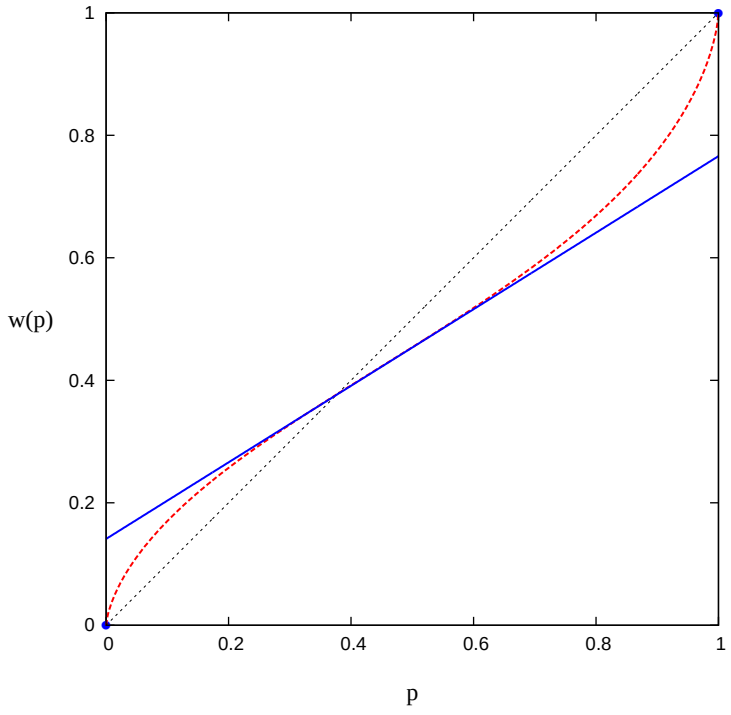
Thus for a PT decision maker whose probability distortion is vanishingly small $\delta \to 1$, the value function curvature over the domain of losses causes the iso-$V^-(C)$ contours to be steeper than the iso-$E(C)$ contours in the classical (risk-neutral) case.

4.3 Neo-additive probability weighting function

The principal shortcomings of the above PT implementation of SDT – namely (i) potential non-uniqueness of the optimal cutoff threshold, and (ii) opaqueness to interpretation as well as manual calculation – may be ameliorated by substitution of a piece-wise linear *neo-additive* probability weighting function for the TK92 probability weighting function. There are numerous precedents for its use.[34–37] Viscusi and Evans[38] present empirical evidence for its use, while Abdellaoui[39,40] provides parameter estimates. Wakker notes that neo-additive weighting functions "are among the most promising candidates regarding the optimal tradeoff of parsimony and fit".[41] They capture both the *possibility* effect, at the transition from impossibility ($p = 0$) to possibility ($p > 0$), as well as the *certainty* effect, at the transition from highly likely to certain ($p = 1$). In-between the two extremes, the linear form overweights small probabilities and underweights large probabilities.

$$w_{\text{n-a}}(p) = \begin{cases} 0 & \text{for } p = 0 \\ \mathsf{a}\,p + \mathsf{b} & \text{for } 0 < p < 1 \\ 1 & \text{for } p = 1 \end{cases} \qquad 0 \le \mathsf{b} < 1, \;\; 0 < \mathsf{a} \le 1 - \mathsf{b} \tag{4.8}$$

Figure 2: Probability weighting functions: TK92 (dashed, red); Neo-additive (solid, blue).



We substitute the neo-additive weighting function (4.8) into (4.5) and solve for the slope of

the iso-$V^{-}_{\text{n-a}}(C)$ contours in ROC space:

$$\frac{\text{dTPL}}{\text{dFPL}} = \left[\frac{(C_{\text{FP}})^{\phi^-} - (C_{\text{TN}})^{\phi^-}}{(C_{\text{FN}})^{\phi^-} - (C_{\text{TP}})^{\phi^-}}\right] \cdot \left(\frac{1-p}{p}\right) \tag{4.9}$$

As is evident from the form of (4.9), the iso-$V^{-}_{\text{n-a}}(C)$ contours are straight lines, just as the iso-$E(C)$ contours of classical SDT. This ensures uniqueness of the optimal cutoff threshold under this criterion when the sampling distributions of θ conditional on $\neg D$ and D respectively are well behaved, yielding everywhere-differentiable ROC curves with monotonically decreasing slope.

The second product term on the rhs in the round brackets is simply the odds of the authentic (appropriate) state $\frac{P(\neg D)}{P(D)}$. This is precisely the manner in which prior probabilities entered the slope of the iso-expected-cost contour expression in classical SDT (3.3). Thus it is not overweighting (underweighting) of small (large) probabilities in itself that is responsible for (i) potential non-uniqueness of the 'optimal' cutoff threshold and (ii) the lack of tractability of the expression for iso-$V^{-}(C)$ contour slope (4.6). Rather, these features are consequences of the non-linearity in the TK92 probability weighting function. From the standpoint of optimal operating point *uniqueness*, the problem is that there is not sufficient regularity in the iso-$V^{-}(C)$ contours to ensure that there will be, *for one and only one* contour, *either* (i) *only one* boundary intersection point, or (ii) *only one* interior tangency point.

The term in square brackets in (4.9) is identical to the corresponding term in (4.6). Hence, as shown in (4.7), the square-bracketed term in (4.9) is larger than the corresponding term in classical SDT. It follows from the properties of the square- and round-bracketed terms together that the iso-$V^{-}_{\text{n-a}}(C)$ contours are steeper than the iso-$E(C)$ contours. Thus PT-SDT under the neo-additive probability weighting function yields a more conservative cutoff threshold than classical SDT.

5 PSYCHOLOGY OF DECEPTION AND SDT

Some deceptive ploys are more likely to succeed than others. The literature reviewed in Section 2 identifies four categories of psychological factors that have been linked to successful deception: peripheral-route persuasion, visceral emotions, urgency, and contextual cues. All other things considered equal, the more skillful the perpetrator and the more effort and resources used to craft a tailored, psychologically and contextually compelling deception ploy, the more likely the ploy will be successful.

Yet in order to progress with model building, it is necessary to concede that all other things

cannot be considered equal. In this vein, we distinguish between (i) the quality of the *match* and (ii) the quality of the *implementation*.

5.1 Match quality

Not every phishing-ploy type – no matter how well implemented – will have equal traction with all users. For instance, an appeal to authority is 100% ineffective on a user who has an antagonistic relationship with authority. Similarly, not all implementation choices will be equally effective on all users. Consider the attempt to trigger the visceral emotion of greed through the device of 'lottery millions'; this will be 100% ineffective on a devout Muslim user, as gambling is *haram* (sinful, forbidden).

Individual phishing emails are comprised of a bundle of cues, some reflecting the chosen phishing-ploy type, others reflecting implementation choices. Although there will therefore be a finite, discrete combination of cues comprising each bundle, we will define the *match quality* as a continuous index between zero and one $m \in [0,1]$.

Let the space of possible cue combinations be represented as $\Gamma = \{0,1\}^z$, where $z \in \mathbb{N}$ is the finite total number of possible cues under consideration. Let $\boldsymbol{\gamma} \in \Gamma$ be the combination of cues contained in the phishing email, where $||\boldsymbol{\gamma}||_0$ is the number cues. In turn, let $\boldsymbol{h}_i \in [0,1]^z$ be user i's z-vector of indices, one for each cue $h_{ij} \in [0,1]$, $j = 1, 2, ..., z$, documenting the quality each potential cue's match with user i. The phishing email's global match quality with user i is therefore given by the function $m : [0,1]^z \times \{0,1\}^z \to [0,1]$, that is

$$m_i \;=\; m(\boldsymbol{h}_i, \boldsymbol{\gamma}) \;\; . \tag{5.1a}$$

For risk simulation purposes it proves unnecessary to fully specify the space of cues Γ and to define match quality as a function of $\boldsymbol{h}_i$ and $\boldsymbol{\gamma}$ as in equation (5.1a). Notice that this (5.1a) approach would require $\boldsymbol{h}_i$ to be defined explicitly for each user and $\boldsymbol{\gamma}$ to be defined explicitly for each phishing email. Also notice that the quality of the correspondence between $\boldsymbol{\gamma}$ and $\boldsymbol{h}_i$ will have a random component because (i) the attacker's knowledge of $\boldsymbol{h}_i$ is incomplete and (ii) the attacker designs the phishing email to be effective across a number of users. In the limiting case of a focused spear-phishing attack targeting a single user, (ii) ceases to be a source of randomness, but (i) still remains. Hence we may abstract from the particularistic details of specific cue sets and their matches with specific individuals, and instead model match quality

as a transformation $m : [0,1] \to [0,1]$ of the standard uniform random variable ξ:

$$m \;=\; m(\xi) \qquad \text{where} \qquad \xi \sim U(0,1) \;\; . \tag{5.1b}$$

This specification relies on the function $m(\cdot)$ to capture the mapping of random draws from $U(0,1)$ to the match quality scale. Three classes of mappings are apposite. First, the class of cumulative distribution functions of continuous, everywhere differentiable probability density functions. A particularly flexible exemplar of this category is the cumulative Beta distribution

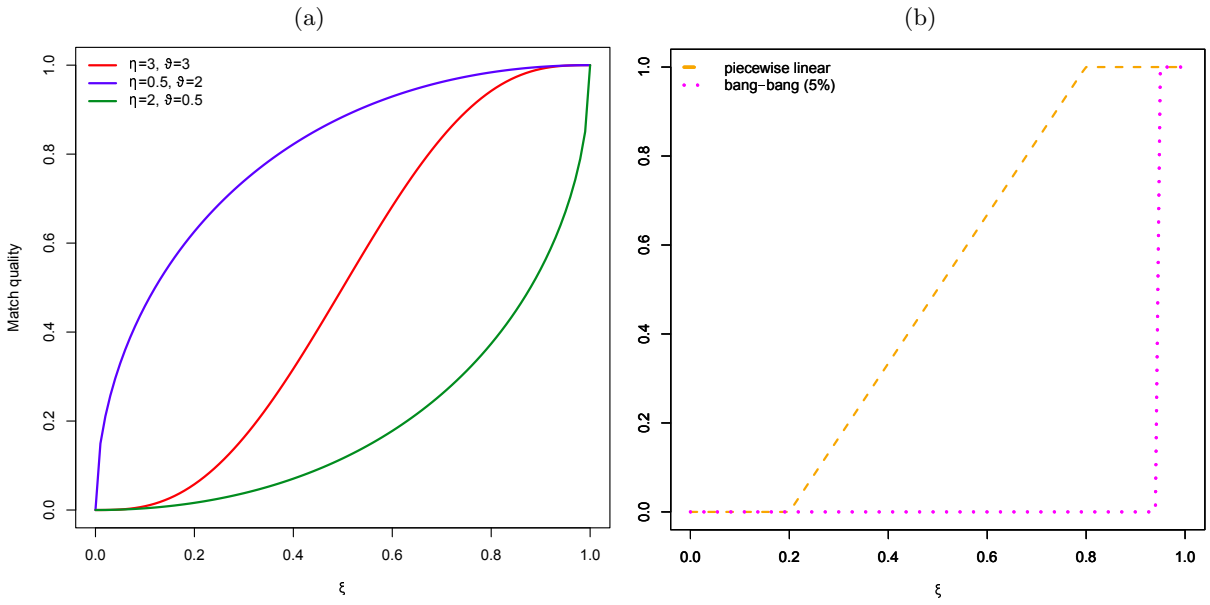
$$m(\xi) = \frac{1}{B(\eta,\vartheta)} \int_0^{\xi} t^{\eta-1}(1-t)^{\vartheta-1}\mathrm{d}t \;\; , \qquad \eta, \vartheta \in \mathbb{R}_{++} \;\; , \tag{5.2}$$

where $B(\eta,\vartheta)$ is the Beta function. The flexibility of the cumulative Beta function mapping is illustrated in Figure (3a). Extreme match quality patterns are also accommodated by this function. For instance with the location parameter held small ($\eta = 0.5$), the cumulative Beta function attains the value of 0.95 for $\xi = 0.038$ when the scale parameter is set to $\vartheta = 50$. Under this mapping, the achieved match quality is very high for all but the lowest values of ξ. With η large and ϑ small, the opposite pattern (i.e. very poor match quality for all but the highest values of ξ) may be represented. And for balanced (η,ϑ) combinations, the relationship between $m(\xi)$ and ξ approaches identity for low values of the location and scale parameters. For high-and-balanced (η,ϑ) combinations, $m(\xi)$ sharply distinguishes between below-average and above-average values of ξ.

The second class of mappings is illustrated with the orange (dashed) line in Figure (3b). This piecewise linear class is straightforward to calculate yet approximates the S-shaped cumulative distribution function form.

Finally, the third class of 'bang-bang' mappings is illustrated with the magenta (dotted) line in Figure (3b). In this particular ($\xi' = 0.95$ threshold) bang-bang mapping, 95% of phishing emails suffer match-quality failure ($m = 0$), while only the top 5% achieve the required match precision ($m = 1$). Notice that the transition threshold's placement is not restricted in principle, i.e. $\xi' \in [0,1]$, but those in the neighborhood of $\xi' = 0.95$ are particularly useful for modeling purposes, as they focus entirely on good-quality matches, which are a small subset of all possible pairings. A model based on a bang-bang mapping with $\xi' = 0.95$ is conservative in the sense that it focuses on the good matches – which also occur in mappings such as those illustrated with the red line in Figure (3a) – but excludes a large number of intermediate-quality matches. This mapping simplifies and clarifies the interpretation of simulation modeling, as will be seen

Figure 3: (a) Match quality mappings from the cumulative Beta distribution. (b) Piecewise linear match quality mappings.



below in Section 6.

5.2 Implementation quality

Under certain assumptions, the AUC of an ROC curve gives a direct reading of the probability that a randomly selected positive will have a higher score value than a randomly selected negative AUC $= P(\theta_D > \theta_{\neg D})$.[24] For normally distributed equal-variance sampling distributions, the AUC is a function of the discrimination parameter d', which is the normalized distance between the means of the two sampling distributions (see equation (3.1)).[4] Denoting the deception perpetrator's skill[5] in deploying and manipulating peripheral-route persuasion, visceral emotion, and contextual cues by $K \in \mathbb{R}_+$, the amount of effort (time, exertion and resources) used in crafting the deception ploy by $e \in \mathbb{R}_+$, and the mark's accumulated experience and learning by $\mathcal{E}_{it} \in \mathbb{R}_+$, then we can express the mark i's discrimation at time t with the differentiable function $d : \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ where

$$d'_{it} = d(K, e, \mathcal{E}_{it}) \quad . \tag{5.3}$$

The time subscript t is necessary to reflect mark i's possibilities for gaining experience and learning to identify particular classes of ploys. The mark's discrimination on a ploy perpetrated

[4]We wish to thank one of the Reviewers for pointing out that the term 'discriminability' is awkward and that its literal meaning may be confusing. Henceforth we diverge from prevailing convention within the SDT literature and refer to 'discrimination' rather than 'discriminability'.

[5]i.e. human capital

at time $t < \tau$, where τ is when the mark learns to dispassionately recognize that ploy type, will be smaller than at any time $t > \tau$.[6]

When $d' > 0$ then AUC$> \frac{1}{2}$, and when $d' < 0$ then AUC$< \frac{1}{2}$. Normally in SDT, ROC curves that do less well than chance (AUC$< \frac{1}{2}$) are not employed in modeling, as it is possible to increase classification success simply by noting that these ROC curves identify non-positives rather than positives, and that those not identified are therefore positives. However, in extremis, deception ploys may be engineered to be very effective. Therefore in principle we may wish to admit all AUC$\in [0, 1]$ and $\frac{\mathrm{dAUC}}{\mathrm{d}d'} > 0$ within this range. But for simplicity we henceforth limit analysis to AUC$\geq \frac{1}{2}$ and $d' \geq 0$. The attacker's first-partial derivatives in (5.3) are non-positive $(K', e' \leq 0)$ and second-partial derivatives are non-negative $(K'', e'' \geq 0)$. The contribution from the mark's experience is opposite, i.e. $\mathcal{E}' \geq 0$, $\mathcal{E}'' \leq 0$. Hence

$$\frac{\mathrm{dAUC}}{\mathrm{d}d'}\frac{\partial d'}{\partial K} \leq 0 \ , \quad \frac{\mathrm{dAUC}}{\mathrm{d}d'}\frac{\partial d'}{\partial e} \leq 0 \ , \quad \frac{\mathrm{dAUC}}{\mathrm{d}d'}\frac{\partial d'}{\partial \mathcal{E}} \geq 0 \ , \qquad \forall\, d' > 0 \ . \tag{5.4}$$

Ideally it would be desirable to express deception-ploy-specific AUC or d' as a function of individual peripheral-route persuasion factors (authority, scarcity, similarity and identification, reciprocation, consistency following commitment, and social proof), urgency, visceral factors (greed, pity, lust, fear, and anxiety), and mark-specific contextual cues. However, it is not well-understood what the natural measures or ordinal indices are for all of these factors, and the – presumably complex and non-linear – interactions between these factors are not fully mapped out in the academic literature. Against this background, (5.3) and (5.4) may be viewed as parsimoniously organizing and presenting the upper effectiveness envelope of these (numerous) interacting deception-ploy elements.

The comparative statics are straightforward. For instance, an individual network user who achieves high discrimination within a particular phishing format will have an ROC curve with more pronounced curvature (and larger AUC) than an individual who has lower discrimination. Consequently, the difference between the classical (3.3) and PT (4.9) optimal trade-offs entails that the magnitude of the bias inherent in incorrectly assuming normative rationality is *larger* for agents with a lower discrimination index, i.e. those with lower ROC curvature and AUC. In other words, PT-SDT shifts the optimal cutoff threshold θ^* and optimal operating point $(\alpha^*, (1 - \beta^*))$ *more* for agents with lower ROC curvature and AUC. Thus the psychology of deception magnifies the effect of behavioral decision making under risk and uncertainty.

[6] Although we track the progress of experience and learning, we do not endogenize learning within the present model, as the primary objective is to investigate the effect of behavioral factors on system-level risk.

5.3 Joint effect

In the absence of psychology-of-deception effects, the user's discrimination parameter is *uncompromised*

$$\overline{d}'_{it} = d(0, 0, \mathcal{E}_{it}) \ . \tag{5.5}$$

When the cues present in a phishing email successfully match the user perfectly, then the user's discrimination parameter is *compromised*

$$\underline{d}'_{it} = d(K, e, \mathcal{E}_{it}) \ . \tag{5.6}$$

Dropping subscripts where possible, we may define the *effective discrimination* parameter d'_e from (5.1b), (5.5) and (5.6) as the match-quality weighted convex combination of the compromised $\underline{d}'$ and uncompromised $\overline{d}'$ discrimination parameters.

$$d'_e \ = \ m \cdot \underline{d}' + (1 - m) \cdot \overline{d}' \tag{5.7}$$

Note that if the cue-match between the phishing email and the user is perfect $(m=1)$, then the user's effective discrimination parameter is simply his compromised discrimination parameter.

6 SYSTEM-LEVEL RISK MODELING

6.1 Approach

The individual-level comparative statics presented in Sections 4 and 5 show the direction that cutoff thresholds shift under the influence of behavioral factors. But these comparative statics do not answer the system-level question: *Are the individual-level behavioral effects quantitatively consequential at the level of the whole network?*

Agent-Based Modeling (ABM) is particularly suited to answering this type of question. Here we implement a demonstration-of-principle intended to quantify the distinction between networks consisting of normatively rational users and networks consisting of behavioral users. Nevertheless it is not solely of academic interest. The probability of a network security breach is the disjunction[7] of the probabilities with which individual users are successfully phished. It is known, however, that disjunctive probabilities are under-weighted in intuitive human probability reasoning.[1,42] Wherever Information Security Officers[8] assess network security risk qualita-

[7] logical OR operator, $\vee$

[8] or holders of the closely related job titles of Information Assurance Officer, Information Security Risk Manager or Security Officer

tively, such under-weighting (bias) is a behavioral possibility. Calibrated variants of the type of modeling apparatus developed here may be used instrumentally by Information Security Officers for unbiased security-breach risk estimation.

6.2 Structure and notation

The basic structure of the model specifies the set of users and their exposure to phishing email. The set of network users is $\mathcal{I}$, with cardinality $I = |\mathcal{I}|$ denoting the number of users, whereby the individual users are indexed as $i \in \{1, 2, ..., I\}$. For present purposes, the number of users is held fixed at $I = 100$. All users are held to be average, receiving 50 emails per working day that pass through the spam filter, totalling 250 emails per working week. We assume that for each user, 1 spear-phishing email passes through the organization's spam filter without detection each week. As a proportion, $\frac{1}{250} = 0.004$ of all emails reaching users' in-boxes are malicious. Weeks are indexed as $t \in \{1, 2, ..., T\}$. We assume that the duration of a spear-phishing attack is three weeks, $T = 3$. In accordance with equation (5.3), a user i who fails to reject a malicious email in week τ will, as a result of the experience and the attendant attention, drastically improve her ability to detect this morphology of phishing email, yielding $\text{AUC}_{it} \approx 1 \ \forall\, t \in \{\tau{+}1, \tau{+}2, ..., T\}$.

After experimentation with the model, an exemplar cost structure was found that allows the properties of the model to be illustrated most clearly. This cost structure – which is not implausible – reflects the network user's experience of the costs of (mis-)classifying spear-phishing emails. For the network user, the consequences of erroneously responding to a spear-phishing email are pre-eminent, by a large margin. Hence we fix $C_{\text{FN}} = 20,\ C_{\text{TP}} = 0.5,\ C_{\text{FP}} = 0.25,\ C_{\text{TN}} = 0$. Furthermore, as a starting point for the benchmark model, we assume that users expect to receive on average one spear-phishing email per day, i.e. a $p = \frac{5}{250} = \frac{1}{50}$ prior probability.[9]

The fraction of behavioral users b whose decision making is best described by PT is an empirical question specific to each organization. But in order to bring the distinction between normative and behavioral into sharp relief, we focus on the extremes: $b \in \{0, 1\}$.

To capture the psychology of deception as described in Sections 2 and 5, we employ a simplified, discretized operationalization that implements the bang-bang match-quality mapping. Let the match quality of the psychological deception ploy at time t with user i be determined by the Bernoulli random variable X_{it}. We specify the users' compromised and uncompromised discrimination parameters as $\underline{d}' \in \mathbb{R}_+$ and $\overline{d}' \in \mathbb{R}_+$ respectively, with $0 \leq \underline{d}' < \overline{d}'$. Conditional on successful cue-match (see Sections 5.1 and 5.3) indicated by $x_{it} = 1$, the discrimination parame-

[9] The consequences of varying this parameter are explored in Section 6.3.2.

ter drops to the low value of $d'_{it} = \underline{d}'$, while conditional upon cue mis-match indicated by $x_{it} = 0$, the discrimination parameter remains at $d'_{it} = \overline{d}'$. Techniques for empirically estimating $\hat{\text{AUC}}$ and therefore $\hat{d}'$ are well established.[23–25] For present purposes we fix $\underline{d}' = 0.5$ (AUC=0.638) and $\overline{d}' = 3.0$ (AUC=0.983). When the attacker fails to achieve high cue-match quality, users' discrimination is high. Probability π can therefore be defined as $\pi := P(x = 1) = P(\underline{d}')$ and $(1 - \pi) := P(x = 0) = P(\overline{d}')$. We fix $\pi = 0.05$. We feel that 5% is a conservative estimate – in the sense of being closer to the lower bound than to the upper bound – of the total population fraction upon which a psychological deception ploy gains at least *some* traction (see e.g. the red and blue match-quality functions in Figure (3a)). Note finally that these (population-level) probabilities are not directly accessible to individual decision makers, and consequently are not endogenized by individual decision makers.

Using these assumptions we conduct three separate anlyses, comparisons between which allow quantification of the network-level consequentiality of individual-level behavioral effects. Firstly, the benchmark scenario of normative rationality (model M0). Secondly, the behavioral decision-making effects codified by PT-SDT (model M1). Thirdly, the PT-SDT behavioral decision-making effects combined with the psychology-of-deception effects on discrimination (model M2). Finally, we run two further analyses to determine the independent effects of varying the network users' prior probability p (Section 6.3.2) and discrimination $\overline{d'}$ (Section 6.3.3).

6.3 Results

6.3.1 Benchmark model

Table 2 summarizes the distribution of security breaches under each model when network users are subjected to a simulated 3-week spear-phishing attack. In this outcome variable the models are ordered in each of their $j \in \{1, 2, 3\}$ quartiles

$$Q_j^{\text{M0}} < Q_j^{\text{M1}} \leq Q_j^{\text{M2}} \qquad \forall j \in \{1, 2, 3\} \tag{6.1}$$

as well as in their means

$$\hat{\mu}^{\text{M0}} < \hat{\mu}^{\text{M1}} < \hat{\mu}^{\text{M2}} \ . \tag{6.2}$$

Figure 4 illustrates the frequency and cumulative distributions of the three models. Inspection indicates a first-order stochastic dominance relationship between model M1 and model M0. Between model M2 and model M1, first-order stochastic dominance also holds down to a 0.1%-

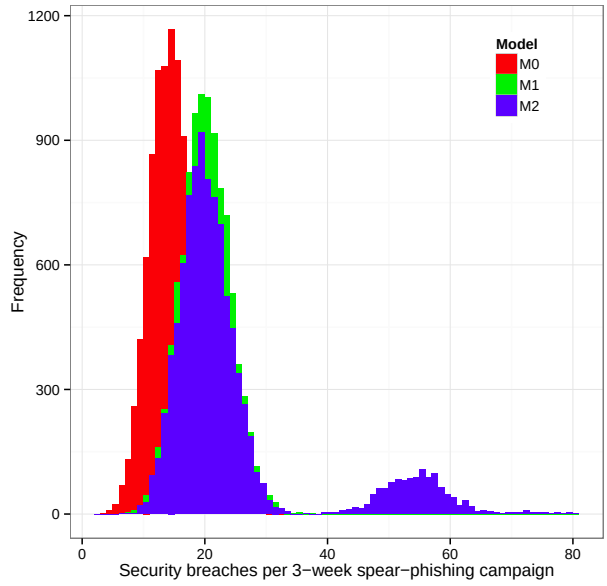
quantile-increment granularity.[10] Considered separately (or alternatively, by transitivity) model M2 first-order stochastically dominates model M0. Note that since the outcome variable is the count of security breaches, the stochastically dominat*ed* model has the lower security risk.

Table 2: Distribution of security breaches in 10,000 repetitions of a 3-week attack.

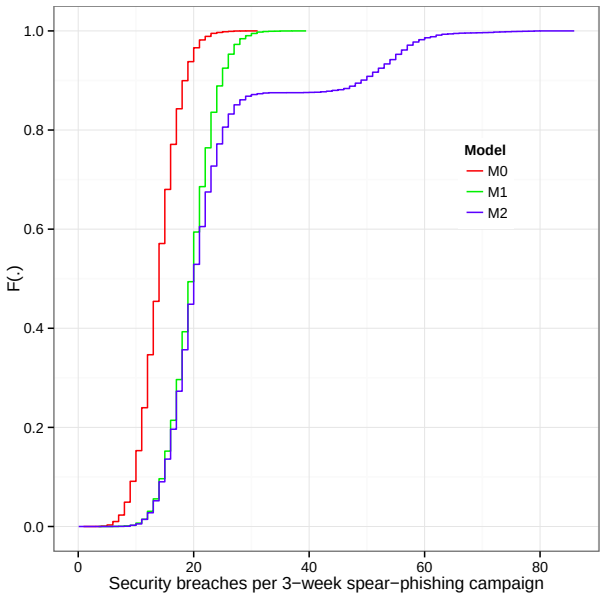
	M0	**M1**	**M2**
Min.	3.0	6.0	6.0
Q_1	12.0	17.0	17.0
Q_2	14.0	20.0	20.0
$\hat{\mu}$	14.0	19.7	23.9
Q_3	16.0	22.0	24.0
Max.	29.0	37.0	80.0

Figure 4: Frequency and cumulative distributions of security breaches under models M0–M2 in 10,000 repetitions of a 3-week spear-phishing campaign.

(a) Frequency distributions



(b) Empirical CDFs



The parameters for this benchmark simulation were chosen in part to reveal the differences between the three models. Accordingly, the mixture distribution generated by model M2 is particularly evident in Figure 4. This reflects the psychology-of-deception effects (drastically weakened discrimination) among a subset of network users, which magnify the behavioral decision

[10] With quantile increments of 0.01%, which register each of the 10,000 observations individually, unrestricted first-order stochastic dominance would hold but for 4 quantiles in the extreme left tail: 0.01%, 0.02%, 0.24% and 0.25%. In their work on statistical testing for stochastic dominance, Davidson and Duclos point out that "...testing for unrestricted dominance is too statistically demanding, since it forces comparisons of dominance curves over areas where there is too little information."[(43)] Hence the emphasis on restricted tests that censor the tails.

making effects already incorporated into PT-SDT. The 5% bang-bang match-quality mapping places a stringent match-quality requirement on the phishing emails, resulting in either fully compromised discrimination (the minority) or completely uncompromised discrimination (the majority). Under any of the other match-quality mappings illustrated in Figure 3, the effective discrimination parameter (5.7) takes intermediate values $d'_e \in (0,1)$ as well, thereby adding density to and filling in the valley between model M2's major and minor modes.

Notice that the minor mode comprises more than 5% of the combined distribution. This is because each attack has a 3-week duration, and those users who did not commit a False Negative misclassification error in week 1 become subjected once more in week 2 to a 5% cue-match success rate psychological deception ploy. And in week 3, those users who did not commit a False Negative misclassification error in week 1 or week 2 are subjected to the 5% cue-match success rate psychological deception ploy for a final time. Since the total number of phishing emails over the three weeks is 300, we can compute the expected fraction of these 300 classification tasks to be conducted under the compromised discrimination parameter $\underline{d}'$. Denoting the proportion of users (out of 100) committing a False Negative misclassification error in week $k \in \{1,2,3\}$ by $\varphi_k \in (0,1)$, the expected share of all classification tasks to be conducted under the compromised discrimination parameter is:

$$\pi(3 - 2\varphi_1 - \varphi_2) \quad . \tag{6.3}$$

The term in brackets will be greater than 1 if $\frac{2}{3} > \varphi_1$ and $\varphi_1 \geq \varphi_2$. So for the parameter combinations employed here, the greater-than-5% share observed in model M2's minor mode is not in itself anomalous.

Empirical Cumulative Distribution Functions (CDFs) as in Subfigure (4b) form the basis for calculating the probability that there will be more than $z \in \{0,1,2,...,I\}$ security breaches. This probability may be computed as $P_{\mathrm{M}j}(\tilde{z} > z) = 1 - \hat{F}_{\mathrm{M}j}(z)$ for $j \in \{0,1,2\}$. Table 3 presents this probability, calculated for each model and for each of five different security-breach-count levels $z = (10, 15, 20, 25, 30)$. Rows four and five report the *bias* – in terms of under-estimated security breach probability – of assuming normatively rational network users instead of PT-SDT behavioral decision makers (row 4) or instead of PT-SDT decision makers subject to psychology-of-deception effects (row 5). These last two rows of Table 3 show that the bias involved in assuming normative rationality is non-trivial.

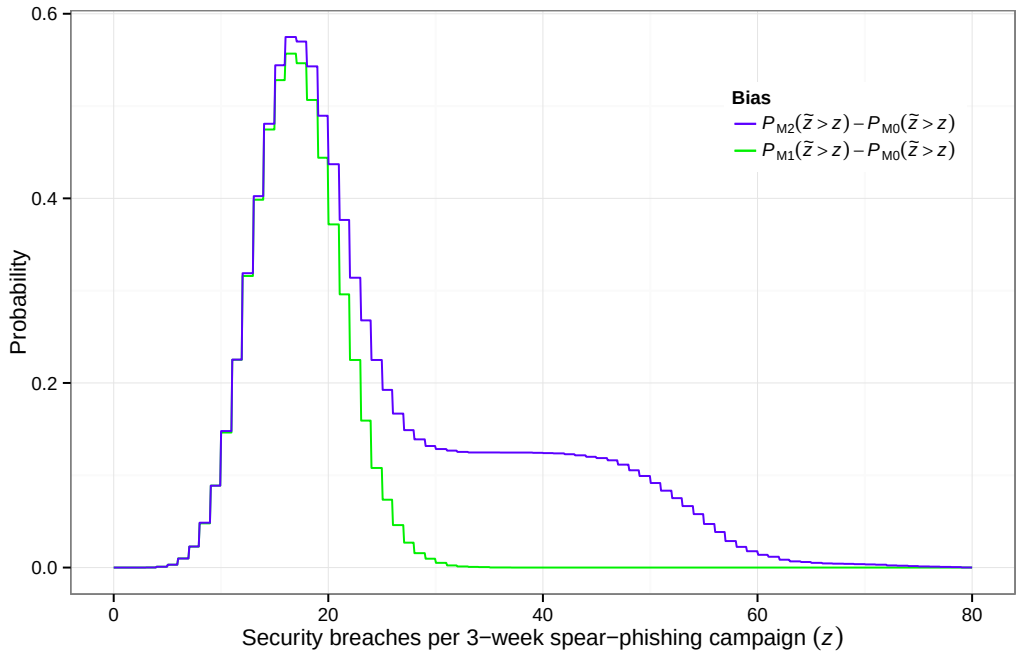
As Figure 5 shows,[11] this under-estimation bias reaches its supremum at $z = 16$. Calculating $P_{\mathrm{M0}}(\tilde{z} > 16)$ when the true model is M1 under-estimates by $P = 0.5568$ the probability that

[11] Note that Figure 5 does not present probability distributions; the area under each curve need not equal unity.

Table 3: Probability that there will be in excess of 10, 15, 20, 25 and 30 security breaches.

		$P.(\tilde{z} > 10)$	$P.(\tilde{z} > 15)$	$P.(\tilde{z} > 20)$	$P.(\tilde{z} > 25)$	$P.(\tilde{z} > 30)$
(1.)	M0	0.85	0.32	0.03	0.002	0.000
(2.)	M1	0.99	0.85	0.41	0.075	0.005
(3.)	M2	0.99	0.86	0.47	0.194	0.128
(4.)	(2.)-(1.)	0.14	0.53	0.38	0.073	0.005
(5.)	(3.)-(1.)	0.14	0.54	0.44	0.192	0.128

Figure 5: Magnitude of under-estimate (bias) in calculating $P_{\text{M0}}(\tilde{z} > z)$ when in fact the descriptively accurate model is M1 (green line) or M2 (blue line).



there will be more than 16 security breaches. Calculating $P_{\text{M0}}(\tilde{z} > 16)$ when the true model is M2 under-estimates by $P = 0.5748$ the probability that there will be more than 16 security breaches. In other words, individual-level behavioral effects are substantial and quantitatively consequential for network-level risk assessment.

Note that this peak bias is also equal to the statistic $D^{-}_{n_o,n_j}$ on samples of size $n_0 = n_j = n$

where $j \in \{1, 2\}$:

$$\begin{aligned} P_{\mathrm{M}j}(\tilde{z} > 16) - P_{\mathrm{M}0}(\tilde{z} > 16) &= (1 - \hat{F}_{\mathrm{M}j}(16)) - (1 - \hat{F}_{\mathrm{M}0}(16)) && (6.4) \\ &= \hat{F}_{\mathrm{M}0}(16) - \hat{F}_{\mathrm{M}j}(16) && (6.5) \\ &= \sup_{z \in \mathcal{I}} \left\{ \hat{F}_{\mathrm{M}0,n_0}(z) - \hat{F}_{\mathrm{M}j,n_j}(z) \right\} && (6.6) \\ &= D^{-}_{n_0,n_j} \ . && (6.7) \end{aligned}$$

The two-sample Kolmogorov-Smirnov test is a test of the null hypothesis that both samples are drawn from the same distribution, and it employs a test statistic that is proportional to $D^{-}_{n_0,n_j}$, the largest positive difference between empirical CDF under normative rationality $\hat{F}_{\mathrm{M}0,n_0}(z)$ and the empirical CDF that reflects behavioral effects $\hat{F}_{\mathrm{M}j,n_j}(z)$. Table 4 presents the difference $D^{-}_{n_j,n_k}$, the K-S test statistic, and associated one-sided p-value in columns 2–4. The hypothesis that samples of size n_j and n_k are drawn from a common distribution is rejected for $j \in \{0, 1\}$, $k \in \{1, 2\}$, $j \neq k$.

Following McFadden,[44] a test for whether $\hat{F}_{\mathrm{M}k}$ first-order stochastically dominates $\hat{F}_{\mathrm{M}j}$ using two independent samples of identical size $n_j = n_k = n$ may be formalized as the test of the null hypothesis $H_0 : \hat{F}_{\mathrm{M}j,n_j}(z) \geq \hat{F}_{\mathrm{M}k,n_k}(z)$ for all z against the alternative of $H_1 : \hat{F}_{\mathrm{M}j,n_j}(z) < \hat{F}_{\mathrm{M}k,n_k}(z)$ for some z, with a significance level given by the probability of rejecting H_0 when $F_{\mathrm{M}j} \equiv F_{\mathrm{M}k}$. McFadden's test statistic is

$$D^{*}_{n} = \sup_{z \in \mathcal{I}} \sqrt{n} \left(\hat{F}_{\mathrm{M}k,n_k}(z) - \hat{F}_{\mathrm{M}j,n_j}(z) \right) \ , \tag{6.8}$$

which has a significance level $P(D^{*}_{n} > q)$ with the large-n limiting distribution of

$$P(D^{*}_{n} > q) \ \backsimeq \ e^{-q^2} \left(1 - \frac{q}{3} \sqrt{\frac{2}{n}} + O(1/n) \right) \ . \tag{6.9}$$

Table 4 presents McFadden's test statistic and its associated p-value in columns 5–6. The null hypothesis that $\hat{F}_{\mathrm{M}k}$ first-order stochastically dominates $\hat{F}_{\mathrm{M}j}$ fails to be rejected in all three instances.

6.3.2 Prior probability

Figure 6 illustrates the relationship between (i) the network user's prior probability of receiving a spear-phishing email and (ii) the distribution of network breaches, ceteris paribus. In the $p = 1/250$ prior probability cell of Subfigure (6a) it is notable that in 26.8% of the 10,000

Table 4: Peak underestimation bias, Kolmogorov-Smirnov (K-S) tests, and First-Order Stochastic Dominance (FOSD) tests.

		Kolmogorov-Smirnov		McFadden FOSD test	
Mj instead of Mk $j\in\{0,1\}$, $k\in\{1,2\}$, $j\neq k$	Bias peak $D^{-}_{n_j,n_k}$	statistic $\sqrt{n}D^{-}_{n_j,n_k}$	one-sided p-value	statistic D^{*}_{n}	one-sided p-value
M0 instead of M1	0.5568	55.68	0.000	0.00	0.999
M0 instead of M2	0.5748	57.48	0.000	0.00	0.999
M1 instead of M2	0.1246	12.46	0.000	0.01	0.999

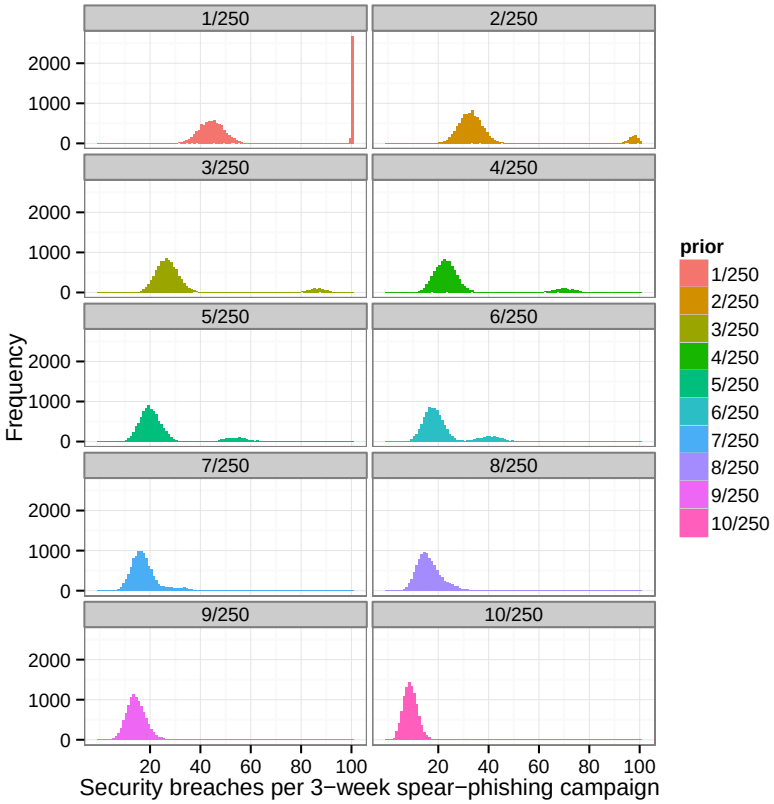
repetitions, all 100 network users fall victim to the 3-week spear-phishing attack. In the opposite extreme cell of Subfigure (6a), which pictures the prior probability $p = 10/250$ case, only 0.01% of the 10,000 repetitions involve 20 or more network users falling victim to the 3-week spear-phishing attack. A prior probability equivalent to expecting on average two spear-phishing emails per weekday (10/250) is sufficient to mitigate the $\pi = 0.05$ psychology-of-deception effect on the distribution of network breaches.

As equation (C.6) in Appendix C shows, the prior probability p affects the optimal cutoff threshold θ^* nonlinearly, via a logarithmic transformation. Replacing $p = 10/250$ with $p = 1/250$ shifts the optimal cutoff threshold to the right by $\frac{2.339}{\mu_D}$, which is $\triangle\theta^* = \frac{2.339}{3} = 0.78$ under $\overline{d'} = 3$ but $\triangle\theta^* = \frac{2.339}{0.5} = 4.68$ under $\underline{d'} = 0.5$. For $d' < 1$ the effect of $\triangle p$ on θ^* is magnified, and $\lim_{d'\to 0} \triangle\theta^* = \infty$. The statistical power of the SDT and PT-SDT classifiers respond as $\lim_{d'\to 0}(1-\beta) = 0$. This is because, as $d' \to 0$, the optimal operating point converges to the boundary $(\alpha^*, (1-\beta^*)) = (0, 0)$ when the slope of the iso-expected-cost lines (3.3) or iso-$V^{-}_{\text{n-a}}(C)$ lines (4.9) exceed unity.

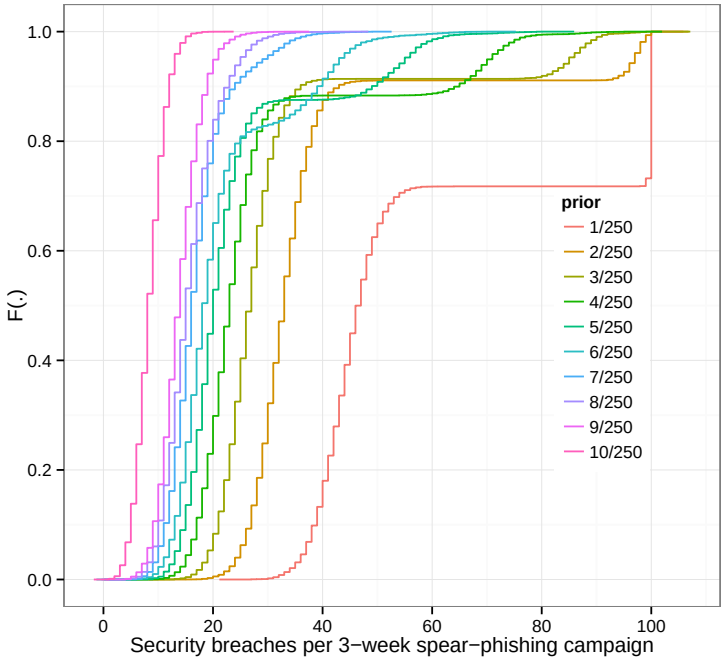
Thus we have elaborated, through numerical analysis, what was found through theoretical analysis in Section 5: that *smaller* d' cause the impact of a change in the optimal operating trade-off – whatever the source of this change – to be manifested in a *larger* shift in the optimal cutoff threshold. The psychology of deception – which for some network users drastically diminishes discrimination – magnifies not only the effect of behavioral decision making under risk and uncertainty, but also the effect of individuals' private beliefs (prior probability) about the likelihood of being targeted by a sophisticated spear-phishing email.

Figure 6: Effect of prior probability (ranging from 1/250 to 10/250) on the frequency and cumulative distributions of security breaches in 10,000 repetitions of a 3-week spear-phishing campaign.

(a) Frequency distributions



(b) Empirical CDFs



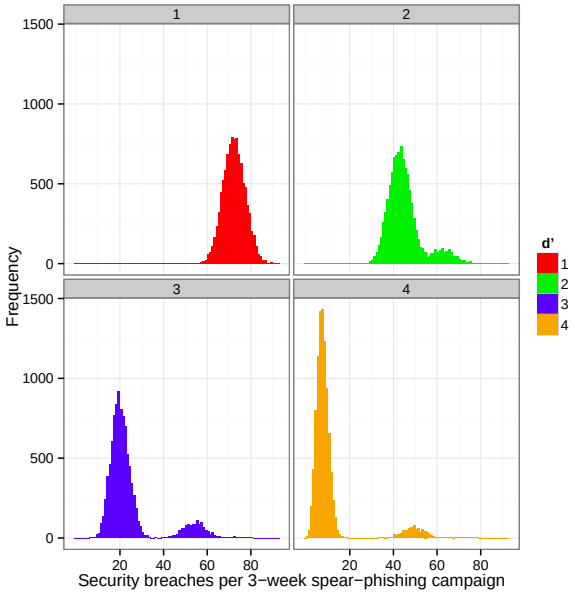
6.3.3 Discrimination

Figure 7 illustrates the effect of varying uncompromised discrimination $\overline{d'}$, ceteris paribus,[12] through the sequence of values $\overline{d'} = (1, 2, 3, 4)$. These values correspond to the AUC values AUC$= (0.760, 0.921, 0.983, 0.998)$. All but the $\overline{d'} = 1$ frequency distributions have a second, minor mode to the right of the major mode. For $\overline{d'} = 4$, truncating the distribution to exclude the minor mode yields a mean of $\hat{\mu}|_{z \leq 18} = 7.02$, which equates to 2.34 security breaches on average per week of the spear-phishing campaign. In turn the minor mode, which comprises 8.37% of all spear-phishing campaigns, has a mean of $\hat{\mu}|_{z > 18} = 49.92$.

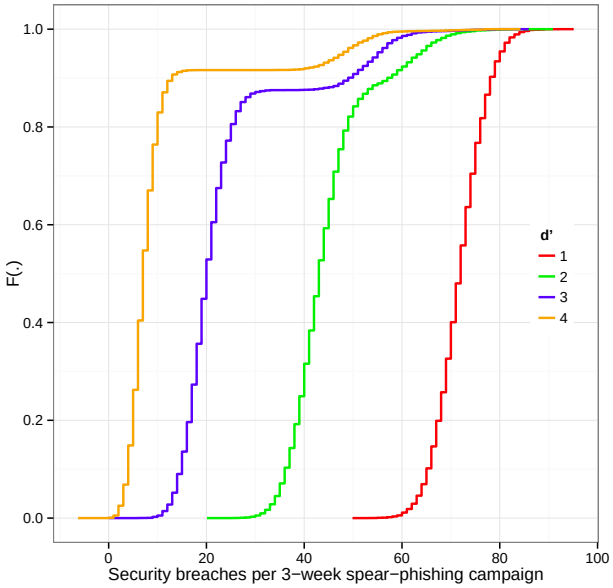
The $\overline{d'} = 1$ cell of Subfigure (7a) reinforces what was seen in the $p = 10/250$ cell of Subfigure (6a): that observing a uni-modal empirical distribution is not sufficient in itself to rule out the existence of an underlying finite mixture structure. As $\overline{d'}$ increases, the major mode shifts to the left. Even though the minor mode depends primarily on $\underline{d'}$, it too is dragged to the left as $\overline{d'}$ increases, though to a lesser extent than the major mode.

Figure 7: Effect of discrimination, ranging from $\overline{d'} = 1$ to $\overline{d'} = 4$, on the frequency and cumulative distributions of security breaches in 10,000 repetitions of a 3-week spear-phishing campaign on model M2 network users.

(a) Frequency distributions



(b) Empirical CDFs



[12] In particular, the compromised discrimination parameter is held at $\underline{d'} = 0.5$.

7 DISCUSSION

7.1 Levers and Implications

Section 6.3.2 demonstrates that the network user's beliefs regarding the frequency of receiving spear-phishing emails is a theoretically effective 'lever' with which to influence the distribution of security breaches. This lever is formalized as the network user's prior probability p that any one email is of the spear-phishing type. In turn Section 6.3.3 demonstrates the effect of the network user's discrimination parameter $\overline{d}'$ on the distribution of security breaches. And controlling the transition from uncompromised discrimination $\overline{d}'$ to compromised discrimination $\underline{d}'$, the parameter π, representing $m(\xi)$, captures the cue-match quality of the psychological deception ploy.

Compared to (mis)classification costs, which tend to be a rather static feature of organizational culture, the parameters p, $\overline{d}'$ and π are in principle responsive to organizational initiatives, training and education.

The model developed in this paper supplies grounds for explicitly including and targeting these parameters in the learning objectives of cybersecurity training programs. With appropriate calibration to a specific organization, the present modeling framework may be used as a tool for determining whether there is a security-risk priority order among p, $\overline{d}'$ and π, and thus for assigning learning intervention priorities.

The frontline technical solution – spam filtering – is more effective against phishing than against spear phishing. However, insofar as the organization's email filtering becomes more effective against spear-phishing emails, the present analysis suggests that there are unintended consequences and that the reduction in security breach risk may not be concomitant with the reduction in unblocked spear-phishing emails. The reason for this is that better network-level spam filtering reduces the network users' direct exposure to phishing emails, and thus reduces network users' perceived prior probability p of any one in-tray email being malicious. As shown in Figure 6, reductions in p have a marked effect on the distribution of security breaches. From the standpoint of an organization's security risk, it is advantageous if individual network users employ an upward-biased prior probability parameter p. This makes it much more likely that network users will indeed detect malicious spear-phishing emails when an Advanced Persistent Threat campaign succeeds in penetrating the organization's spam filter. The operation of spam filtering could potentially be modified so as to maintain existing levels of technical protection but without diminishing users' awareness of the frequency with which the the organization is targeted with malicious email. Potential measures to achieve this range from weekly status reports on

phishing and spear-phishing emails to more elaborate systems for capturing malicious emails, selecting a (small) subset of these emails, substituting their malicious links and attachments with benign content, followed by releasing the emails to their original intended recipients. Recipients of these trapped-cleansed-and-released emails either detect their true nature and delete and/or report them, or they succumb to the (now harmless) deception ploy, by which they themselves and the Network Security Officer are alerted of the need to increase p, revise personal email procedures, and possibly attend additional training.

7.2 Reference points and gains

Given that the analysis in this paper is predicated on the (mis)classification cost structure in equation (4.1), it is apt to query whether and how the results would change if it were possible to associate some benefit or gain with specific (mis)classifications. But the issue of whether outcomes are coded as gains or as losses reflects more fundamental assumptions concerning reference-point determination.

Reference-point determination is the subject of active theoretical and empirical investigation. Within the PT literature, a variety of different non-stochastic reference points have been employed: status quo, lagged status quo, mean of the lottery, and the certainty equivalent. The most recent, third-generation PT has been developed with a stochastic reference point.[47] And in a related literature that eschews probability weighting, Köszegi and Rabin have developed a reference point that arises from endogenously determined lagged beliefs.[45,46] Where the reference point does not arise naturally from the context, there is not widespread consensus on which reference-point concept is apposite and robust.

An arguably natural way of approaching reference-point determination in the phishing context begins with the observation that the user ultimately has only two actions to choose from: Read and Don't Read. Each action defines its own reference point. Where the user decides to Read, (i) the payoff under $\neg D$ is the reference-point payoff of zero, while (ii) the payoff under D is the loss $L_R \in \mathbb{R}_+$. Where the user opts for Don't Read, (iii) the payoff under D is the reference-point payoff of zero, while (iv) the payoff under $\neg D$ is the loss $L_{DR} \in \mathbb{R}_+$ ($L_R > L_{DR}$). Using the editing rule of 'combining' (a.k.a. coalescing), we may write the PT value of each action as:

$$V^-(\text{Read}) = -\lambda w^-(p\beta_{\theta'})v^-(L_R) - \lambda[1 - w^-(p\beta_{\theta'})]v^-(0) \tag{7.1}$$

$$V^-(\text{Don't Read}) = -\lambda w^-((1-p)\alpha_{\theta'})v^-(L_{DR}) - \lambda[1 - w^-((1-p)\alpha_{\theta'})]v^-(0) \tag{7.2}$$

Then the optimal cutoff threshold θ^* can be found as the θ' that equalizes the PT value of the READ option with the PT value of the DON'T READ option.

$$w^-(p\beta_{\theta^*})v^-(L_R) = w^-((1-p)\alpha_{\theta^*})v^-(L_{DR}) \ , \qquad \frac{\partial\,\alpha}{\partial\,\theta'} < 0 \ , \ \frac{\partial\,\beta}{\partial\,\theta'} > 0 \ . \tag{7.3}$$

Then the user's optimal decision is to READ if $\theta \leq \theta^*$ and DON'T READ if $\theta > \theta^*$. Despite the altered approach to determining reference points, payoffs remain within the domain of losses.

Richard Thaler introduced the notion that small gains kept separate from larger losses provide a *silver lining* that is absent if the gains are netted with the losses.[48] In turn Jarnebrant et al. show that, under the PT value function, "segregating a small gain from a larger loss results in greater psychological value than does integrating them into a smaller loss."[49] So potentially there may be scope for network administrators to institute policies that bring about such segregation of small gains from losses/costs, with the objective of guiding – technically, 'nudging' – users' email-classification behavior.

Notice that, holding other factors constant in (7.3), the θ' must move to the left, making $\beta_{\theta'}$ smaller and $\alpha_{\theta'}$ bigger, in order to compensate for $L_R > L_{DR}$. Potentially, one could segregate a small gain G_{DR} such that $L'_{DR} - G_{DR} = L_{DR}$. Then, under DON'T READ when $\neg D$ obtains, rather than incurring the single (loss) payoff of L_{DR}, the user receives the mixed payoff L'_{DR} *and* G_{DR}. According to the silver lining effect, this has a less extreme PT value than L_{DR} alone, entailing additional leftward-shifting of θ' before equality between V^-(READ) and V^-(DON'T READ) is reached.

This brief analysis suggests that a scheme to offer users a partial reimbursement of the delay and inconvenience incurred in diverting falsely suspected emails through a robust screening system may have the desired effect of reducing reporting thresholds. Notice that unlike analyses based on normative rationality – which recommend increased penalties for False Negatives and decreased penalties for False Positives – here the manipulation of payoffs introduces no net change, is focused on False Positives alone, and has an entirely behavioral transmission mechanism.

7.3 Action points

The foregoing suggests action points for training and administration policy as well as for further study and investigation.

Training is suggested for all three variables identified as key levers: the prior probability p, discrimination $d(\cdot, \cdot, \mathcal{E}_{it})$, and the match-quality mapping m. Unlike training to set the prior

probability p at an effective level, training to increase discrimination $d(\cdot, \cdot, \mathcal{E}_{it})$ and to restrict match quality m is likely to suffer from attackers' ability to 'move the goalposts'. Given the scope that attackers have for changing the nature of phishing emails, training may be consigned to a trailing position. However, this simply means that training needs to be flexible, responsive, and delivered in frequency-quantity combinations that are fit for purpose.

Two administration-policy action points are noteworthy. First, selective release of messages from the trapped stream of phishing emails, suitably cleansed of malicious content, may be introduced to stoke users' alertness and to increase their prior probability parameter. Second, nudge policies may be developed which exploit particular features of the behavioral nature of users' decision making. For instance the silver lining effect, applied to False Positives, is expected to lower the cutoff threshold θ^* and to increase detection and reporting rates.

Action points for further study and investigation revolve around calibration of the model. As a first follow-on step, calibration will allow quantification of the sizes of the various effects identified in the present work. In turn, empirical investigation may determine how users update their prior probability estimates with differing exposure rates to 'cleansed' phishing emails. With this in hand, the optimal exposure rate may be determined.

8 CONCLUSION

This work reprises the SDT framework for the purpose of bottom-up modeling of system-level risk. Whereas previous behavioral work has offered general insights motivated by largely verbal analysis, the present results are predicated on integrated mathematical modeling of the consequences of PT-based behavioral decision making. Cognate work on deception, trust and detection has increased our understanding of the psychology of phishing[7–10] – and this understanding underpins improved technical and procedural protocols – but it leaves open the question of how to model the individual-level decision making in a manner compatible with the requirements for system-level risk modeling.

This paper shows that, once augmented with a behavioral PT-based objective function, SDT can viably fulfil this role. In turn the findings of the psychological work on deception, trust and detection may naturally be interpreted in terms of the discrimination parameter d' of SDT, which determines the curvature of the ROC curve and its coverage of the unit square, i.e. the Area Under the Curve (AUC). As proof of the concept, all of these effects have been modeled not only mathematically for a single decision maker, but also numerically at the system level through an Agent-Based Model implementation.

Going forward, one may model and represent the matching between detection skills and specific phishing formats with these analytical concepts, including changes in the quality of the match due to (a) strategic manipulation of the phishing email's properties by the attacker, and, on the part of the users, (b) detection training and learning to influence discrimination and (c) emotion regulation skills training to influence the effectiveness of psychology-of-deception ploys.

Furthermore, as SDT is explicitly built around the matrix of (mis)classification costs, it ensures consistent treatment of all four of these cost items. Using the PT-SDT formulation with neo-additive probability weighting function, it is also straightforward to evaluate the effects of changes in incentives – via the (mis)classification cost matrix – on the optimal cutoff thresholds employed by behavioral lay agents, with a view to aggregating this up to system-level risk.

The present work demonstrates the materiality of incorporating individual-level behavioral biases into the analysis of system-level risk, for example network security risk.[(4)] The PT-SDT model with neo-additive probability weighting function is a tractable means of doing so.

References

1. Kahneman D, Tversky A. (eds). Choices, Values, and Frames. New York, NY: Cambridge University Press, 2000.

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

3. West R. The psychology of security; Why do good users make bad decisions? Communications of the ACM, 2008; 51:34–40.

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

5. Herley C. So long, and no thanks for the externalities: The rational rejection of security advice by users. Proceedings of the New Security Paradigms Workshop (Oxford, UK, Sept 8–11, 2009). NSPW '09. New York, NY: ACM, 2009: 133–144.

6. Jakobsson M, Myers S. Phishing and Countermeasures. New York, NY: Wiley, 2007.

7. Grazioli S, Järvenpää SL. Perils of internet fraud: An empirical investigation of deception and trust with experienced internet consumers. IEEE Transactions on Systems, Man and Cybernetics—Part A: Systems and Humans, 2000; 30:395–410.

8. Johnson PE, Grazioli S, Jamal K, Berryman RG. Detecting deception: Adversarial problem solving in a low base-rate world. Cognitive Science, 2001; 25:355–392.

9. Grazioli S. Where did they go wrong? An analysis of the failure of knowledgeable internet consumers to detect deception over the internet. Group Decision and Negotiation, 2004; 13:149–172.

10. Wright R, Chakraborty S, Basoglu A, Marett K. Where did they go right? Understanding the deception in phishing communications. Group Decision and Negotiation, 2010; 19:391–416.

11. Acquisti A, Grossklags J. Losses, gains, and hyperbolic discounting: An experimental approach to information security attitudes and behavior. In: Camp LJ, Lewis S (eds). The Economics of Information Security. Boston, MA: Kluwer Academic Publishers, 2004:165–178.

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

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

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

15. 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.

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

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

18. 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.

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

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

22. Jagatic TN, Johnson NA, Jakobsson M, Menczer F. Social phishing. Communications of the ACM, 2007; 50:94–100.

23. Egan JE. Signal Detection Theory and ROC Analysis. London: Academic Press, 1975.

24. Green DM, Swets JA, Signal Detection Theory and Psychophysics. London: Wiley, 1966.

25. Macmillan NA, Creelman CD. Detection Theory: A User's Guide. Cambridge: Cambridge University Press, 1991.

26. Ulehla ZJ. Optimality of perceptual decision criteria. Journal of Experimental Psychology, 1966; 71:564–569.

27. Galanter E. Psychological decision mechanisms and perception. In: Carterette EC, Friedman MP (eds). Handbook of Perception II: Psychophysical Judgement and Measurement. New York, NY: Academic Press, 1974:85–126.

28. Healy AF, Kubovy M. Probability matching and the formation of conservative decision rules in a numerical analog of signal detection. Journal of Experimental Psychology: Human Learning and Memory, 1981; 7:344–354.

29. Shermer M, Wheatgrass juice and folk medicine: Why subjective anecdotes often trump objective data. Scientific American, 2008; 299:42.

30. Brandstätter E, Gigerenzer G, Hertwig R. The priority heuristic: Making choices without trade-offs. Psychological Review, 2006; 113:409–432.

31. Glöckner A, Betsch T. Do people make decisions under risk based on ignorance? An empirical test of the priority heuristic against cumulative prospect theory. Organizational Behavior and Humand Decision Processes, 2008; 107:75–95.

32. Glöckner A, Pachur T. Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory. Cognition, 2012; 123:21–32.

39. Abdellaoui M. Parameter-free elicitation of utility and probability weighting functions. Management Science, 2000; 46:1497–1512.

34. Bell DE. Disappointment in decision making under uncertainty. Operations Research, 1985; 33:1–27.

35. Cohen M. Security level, potential level, expected utility: A three-criteria decision model under risk. Theory and Decision, 1992; 33:101–134.

36. Chateauneuf A, Eichberger J, Grant S. Choice under uncertainty with the best and worst in mind: Neo-additive capacities. Journal Economic Theory, 2007; 137:538–567.

37. Abdellaoui M, L'Haridon O, Zank H. Separating curvature and elevation: A parametric probability weighting function. Journal of Risk and Uncertainty, 2010; 41:39–65.

38. Viscusi WK, Evans WN. Behavioral probabilities. Journal of Risk and Uncertainty, 2006; 32:5–15.

39. Abdellaoui M. Parameter-free elicitation of utility and probability weighting functions. Management Science, 2000; 46:1497–1512.

40. Abdellaoui M, Vossmann F, Weber M. Choice-based elicitation and decomposition of decision weights for gains and losses under uncertainty. Management Science, 2005; 51:1384–1399.

41. Wakker P. Prospect Theory for Risk and Ambiguity. Cambridge: Cambridge University Press, 2010.

42. Bar-Hillel M. On the subjective probability of compound events. Organizational Behavior and Human Performance, 1973; 9:396–406.

43. Davidson R, Duclos J-Y. Testing for restricted stochastic dominance. Econometric Reviews, 2012; 32:84–125.

44. McFadden D. Testing for stochastic dominance. In: Romby TB, Seo TK (eds). Studies in the Economics of Uncertainty in Honor of Josef Hadar. New York, NY: Springer-Verlag.

45. Köszegi B, Rabin M. A model of reference-dependent preferences. Quarterly Journal of Economics, 2006; 121:1133–1165.

46. Köszegi B, Rabin M. Reference-dependent risk attitudes. American Economic Review, 2007; 97:1047–1073.

47. Schmidt U, Starmer C, Sugden R. Third-generation prospect theory. Journal of Risk and Uncertainty, 2008; 36:203–223.

48. Thaler RH. Mental accounting and consumer choice. Marketing Science, 1985; 4:199–214.

49. Jarnebrant P, Toubia O, Johnson E. The silver lining effect: Formal analysis and experiments. Management Science, 2009; 55:1832–1841.

APPENDICES

A Probability (weighted) term in the iso-$V^{-}(C)$ contour slope expression

Here is the full form of the probability (TK92 weighted) term appearing in (4.6), $\left(\frac{\psi_1(\alpha,\beta|p,\delta)}{\psi_2(\alpha,\beta|p,\delta)}\right)$:

$$\begin{aligned}\psi_1(\alpha,\beta|p,\delta) = \\ &(1+p(-1+(1-\beta)))(-1+(1-\beta))\left((1+p(-1+(1-\beta)))^{\delta}+(p-p(1-\beta))^{\delta}\right)^{\frac{1+\delta}{\delta}} \\ &\times\Big((-1+p)(-1+\alpha)(p+\alpha-p\alpha)^{2\delta}(-1+\delta) \\ &\quad+((-1+p)(-1+\alpha)(p+\alpha-p\alpha))^{\delta}(p+\alpha-p\alpha+(-1+p)(-1+\alpha)\delta)\Big)\end{aligned} \tag{A.1}$$

and

$$\begin{aligned}\psi_2(\alpha,\beta|p,\delta) = \\ &\left((p-p(1-\beta))^{\delta}(-1+\alpha)(-p+(-1+p)\alpha)\left(((-1+p)(-1+\alpha))^{\delta}+(p+\alpha-p\alpha)^{\delta}\right)^{\frac{1+\delta}{\delta}}\right) \\ &\times\Big((p-p(1-\beta))^{\delta}-\left((1+p(-1+(1-\beta)))^{\delta}+(p-p(1-\beta))^{\delta}\right) \\ &\quad\times(p(-1+(1-\beta))(-1+\delta)+\delta)\Big)\end{aligned} \tag{A.2}$$

B Derivation with neo-additive probability weighting function

Substitute the neo-additive probability weighting function $w_{\text{n-a}}(p)$ from (4.8) into equation (4.5)

$$\begin{aligned}V^{-}(C) \;=\; &-\;[\mathsf{a}p\beta+\mathsf{b}]\lambda[v^{-}(C_{\text{FN}})-v^{-}(C_{\text{TP}})] \\ &-\;[\mathsf{a}p+\mathsf{b}]\lambda[v^{-}(C_{\text{TP}})-v^{-}(C_{\text{FP}})] \\ &-\;[\mathsf{a}(p+(1-p)\alpha)+\mathsf{b}]\lambda[v^{-}(C_{\text{FP}})-v^{-}(C_{\text{TN}})] \\ &-\;\lambda v^{-}(C_{\text{TN}})\;\;.\end{aligned} \tag{B.1}$$

The total differential of this expression, set to zero:

$$\mathsf{a}p\lambda[v^{-}(C_{\text{FN}})-v^{-}(C_{\text{TP}})]\text{dTPL}-\mathsf{a}(1-p)\lambda[v^{-}(C_{\text{FP}})-v^{-}(C_{\text{TN}})]\text{dFPL}=0 \tag{B.2}$$

from which λ and a cancel out, giving the slope of the iso-$V^{-}_{\text{n-a}}(C)$ contours as

$$\frac{\text{dTPL}}{\text{dFPL}}=\left[\frac{v^{-}(C_{\text{FP}})-v^{-}(C_{\text{TN}})}{v^{-}(C_{\text{FN}})-v^{-}(C_{\text{TP}})}\right]\cdot\left(\frac{1-p}{p}\right) \tag{B.3}$$

consistent with (4.9).

C Mathematical ingredients for the system-level model

The system-level aggregation model takes different discrimination parameter $d' \geq 0$ values as inputs, combines these with misclassification cost matrix values, and computes the associated optimal cutoff threshold θ^*, from which the associated optimal true positive $(1-\beta^*)$ and false positive α^* likelihoods are computed, which in turn are used repeatedly in the Agent-Based Model by each agent.

For score values θ drawn from normal sampling distributions for $\neg D$ and D, the true positive and false positive likelihoods may be written in terms of the standard normal Cumulative Distribution Function

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}z^2} d\, z \tag{C.1}$$

for arbitrary cutoff thresholds θ' as:

$$(1-\beta) = \Phi\left(\frac{\mu_D - \theta'}{\sigma_D}\right) \quad , \quad \alpha = \Phi\left(\frac{\mu_{\neg D} - \theta'}{\sigma_{\neg D}}\right) \quad . \tag{C.2}$$

In this bi-normal case, the AUC has the form

$$\text{AUC} = \Phi\left(\frac{\mu_D - \mu_{\neg D}}{\sqrt{\sigma_D^2 + \sigma_{\neg D}^2}}\right) \quad , \tag{C.3}$$

and therefore with identical unit standard deviations $\sigma_D = \sigma_{\neg D} = 1$, AUC$= \Phi\left(\frac{d'}{\sqrt{2}}\right)$.

Maintaining the identical unit standard deviation assumption, the slope of the bi-normal ROC curve is

$$\frac{\mathrm{d}(1-\beta)}{\mathrm{d}\alpha} \;=\; \exp\left\{\left(\frac{\theta' - \mu_{\neg D}}{\sigma} - \frac{d'}{2}\right) d'\right\} \quad . \tag{C.4}$$

The right-hand side of (C.4) may be equated with the slope of iso-expected-cost lines (3.3) for normative decision makers or the slope of iso-$V^-_{\text{n-a}}(C)$ lines (4.9) for behavioral decision makers, in order to solve for the respective optimal cutoff threshold θ^*. Noting further that we may set $\mu_{\neg D} = 0$ without loss of generality, it follows that $d' = \mu_D$ and that for normative decision makers

$$\theta^* \;=\; \frac{1}{\mu_D}\left(\ln(C_{\text{FP}} - C_{\text{TN}}) - \ln(C_{\text{FN}} - C_{\text{TP}}) + \ln(1-p) - \ln p + \frac{\mu_D^2}{2}\right) \quad . \tag{C.5}$$

This cutoff threshold may be substituted back into (C.2) to obtain $(1-\beta^*)$ and α^*. For PT decision makers, the corresponding equation is

$$\theta^* \;=\; \frac{1}{\mu_D}\left(\ln[\,(C_{\text{FP}})^{\phi^-} - (C_{\text{TN}})^{\phi^-}] - \ln[\,(C_{\text{FN}})^{\phi^-} - (C_{\text{TP}})^{\phi^-}] + \ln(1-p) - \ln p + \frac{\mu_D^2}{2}\right) \quad . \tag{C.6}$$