Modeling Phishing Susceptibility as Decisions from Experience

Edward A. Cranford (cranford@cmu.edu)
Department of Psychology, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15221 USA

Kuldeep Singh (kuldeep2@andrew.cmu.edu)
Department of Social and Decision Sciences, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15221 USA

Palvi Aggarwal (palvia@andrew.cmu.edu)
Department of Social and Decision Sciences, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15221 USA

Christian Lebiere (cl@cmu.edu)
Department of Psychology, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15221 USA

Cleotilde Gonzalez (coty@cmu.edu)
Department of Social and Decision Sciences, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15221 USA

Abstract

Traditional anti-phishing training is often non-personalized and does not typically account for human experiential learning. However, to personalize training, one requires accurate models and predictions of individual susceptibility to phishing emails. The present research is a step toward this goal. We propose an Instance-Based Learning model of phishing detection decision-making, constructed in the ACT-R cognitive architecture. We demonstrate the model's ability to predict behavior in a frequency training study, and its generality by predicting behavior in another phishing detection study. The results shed additional light on human susceptibility to phishing emails and highlight the effectiveness of modeling phishing detection as decisions from experience. We discuss the implications of these results for personalized anti-phishing training.

Keywords: phishing; cybersecurity; personalized training; decision making; instance-based learning theory; ACT-R

Introduction

Despite significant advances in security technologies, a large number of phishing emails continue to evade automated detection and are often successful because it is cognitively challenging for humans to distinguish the rare deceptive phishing message from benign emails. As such, phishing attacks remain the biggest, growing threat for cybersecurity (APWG Phishing report, 2020). While phishing attacks exploit human weaknesses using social engineering and psychological techniques (Jagatic et al., 2007), defenders typically employ technological solutions to defend against them, such as machine learning filtering of phishing emails, email authentication tools, and URL filtration/blacklisting (Prakash et al., 2010; Marchal et al., 2014; Peng, Harris, & Sawa, 2018). However, attackers are persistent and phishing emails continue to reach their victims. Since the success of phishing attacks relies on exploiting cognitive and psychological weaknesses, it becomes essential to understand the underlying decision-making processes that influence end-user susceptibility to phishing emails (Canfield, Fischhoff, & Davis, 2016).

Recent research has shown that end-user phishing detection decisions are similar to other kinds of decisions from experience (e.g., Hakim et al., 2020; Singh et al., 2019, 2020). An individual's personal history and experience with emails can have a large influence on phishing susceptibility. Specifically, phishing decisions are influenced by the recency, frequency, and similarity of past emails to the features of the current email. For example, Singh et al. (2019) manipulated the frequency of phishing emails in an anti-phishing training study. The results showed that increasing the frequency of phishing emails during training increased the hit rate of detecting phishing emails. In other research, Singh et al. (2020) examined how the similarity of email features influenced detection accuracy. Their results showed that detection accuracy suffered the more similar the features of a phishing email were to the features of the benign emails. Lastly, Hakim et al. (2020) developed a regression model of end-user phishing susceptibility in an email rating task and revealed an effect of recency on detection decisions. In the task, end-users rated phishing and ham emails on a scale of suspiciousness. The regression model showed evidence of sequential effects of the emails, such that current ratings were positively affected by the previous rating.

Although the evidence shows that phishing decisions are influenced by experiential learning with emails, current training procedures do not take these factors into account, nor have the effects been investigated. Organizations typically

use embedded-training methods that involve sending simulated phishing emails and only provide more traditional phishing training whenever one clicks on the link in the simulated phishing message (Kumaraguru et al., 2009; Kumaraguru et al., 2007). Traditional techniques have often focused on teaching end-users to understand and identify the relevant features that distinguish phishing emails from benign ones (Kumaraguru et al., 2009; Singh et al., 2020). However, there is a deficit of effective experiential phishing training methods that directly address important underlying human cognitive processes in context. Traditional phishing training is often generic and non-personalized. That is, all end-users receive the same set of training emails, the non-phishing emails lack the familiar context that personal ham emails tend to have (e.g., from senders who are familiar to the end-user), and the phishing emails are sent without consideration of the individual's history. Consequently, in many phishing-detection tasks, end-users have trouble distinguishing the phishing emails from the ham, and due to the generic nature of training, the effects of training vary considerably between individuals. In addition, different types of phishing emails have had varied effectiveness across individuals, further emphasizing the need to personalize anti-phishing training (Lin et al., 2019; Oliveira et al., 2017).

Personalized training interventions could prove immensely useful for improving anti-phishing detection, but such methods require models that can be tailored to individuals and that can make accurate decision predictions for a specific phishing email presented at a specific time. Therefore, as a first step toward this goal, we propose a cognitive model that leverages the influence of individual experience on phishing detection decisions, specifically turning to a memory-based theory of experiential learning called instance-based learning theory (IBLT; Gonzalez, Lerch, & Lebiere, 2003). According to IBLT, decisions are made by generalizing across past experiences, or instances, that are similar to the current situation. Typically, instances represent the features of the decision, the action taken, and the outcome of that decision. However, for emails, there is usually a dissociation between the actions taken and feedback regarding whether the email was ultimately malicious. For a given email, IBLT suggests that end-users make decisions by retrieving a classification from memory based on the similarity of features of the current email to features of past emails. Thus, decisions are influenced by typical memory effects such as recency and frequency of past instances and are susceptible to cognitive biases that emerge from these memory processes (e.g., confirmation bias; Lebiere et al., 2013).

General cognitive theories of decisions from experience indicate that the low frequency of phishing emails (compared to benign emails) could be a major issue in the success of detection decisions if end-users underweight the probability of these rare events (Gonzalez et al., 2003; Gonzalez & Dutt, 2011). Additionally, phishing emails often mimic quite well the benign (i.e., ham) emails that regularly flood our inboxes. In other words, phishing emails are similar to the highly frequent and usually recent benign emails that we receive regularly, and phishing decisions are susceptible to effects of frequency, recency, and similarity of features.

Our cognitive model builds upon that proposed by Cranford et al. (2019). In this paper, we first extend and improve upon that model to explore the effects of frequency on phishing detection training by modeling the Phishing Training Task (PTT) in Singh et al. (2019; 2020). We then demonstrate the model's generality by running it through the task in Hakim et al. (2019), the Phishing Email Suspicion Test (PEST), which tests on a different database of emails. Finally, we discuss the implications of the model for future research towards personalized, adaptive anti-phishing training interventions.

Modeling the Phishing Training Task

The PTT (Singh et al., 2019) was designed to examine the impact of learning factors (e.g., frequency effects) on phishing detection decisions. The task is based on the design in Canfield et al. (2016) in which participants are presented a series of email messages and are requested to make classification decisions. In the PTT, participants make three responses to each email: a classification decision of whether the email was a phishing email or not, a confidence rating of their decision (from 50, "not confident at all", to 100, "fully confident"), and the action they would take in response to each email (selected from a 6-point, Likert-type scale ranging from "Respond to this email" to "Report this Email"). For the present model, we focused on the first classification decision.

The PTT consists of three phases: pre-test, training, and post-test. During the pre- and post-test phases, end-users are presented with 10 emails, two of which are phishing emails, and the remaining are benign, ham emails. During the training phase, end-users are presented 40 emails of which 10 20, or 30 are phishing emails. End-users are randomly assigned to one of the three phishing frequency conditions. Feedback about decision accuracy is provided after each trial during the training phase but not during either testing phase.

IBL Model Description

The IBL model was adapted from Cranford et al. (2019) and constructed in the ACT-R cognitive architecture (Anderson & Lebiere, 1998). The modifications made to the model were few, but important, and provided substantial improvement to predicting human behavior in the PTT. These will be discussed below, after presenting the model results.

The model performs the PTT in the same way as humans, processing one email at a time, judging whether each is phishing or ham. For each email, the model takes the content of the email as input and generates a classification by retrieving from similar past instances. For the PTT, the elements of an email include the sender's email address, subject line, email body, link text, and underlying link URL. The classification (i.e., decision) is either phishing or ham. In ACT-R, the retrieval of past instances is based on the activation strength of the relevant chunk in memory and its similarity to each of the elements of the current situation. The activation A_i of a chunk i is determined by the following equation:

$$A_i = \ln \sum_{j=1}^{n} t_j^{-d} + MP * \sum_{k} Sim(v_k, c_k) + \varepsilon_i \quad (1)$$

The first term provides the power law of practice and forgetting, where t_j is the time since the jth occurrence of chunk i and d is the decay rate of each occurrence. The second term reflects a partial matching process, where $Sim(v_k,c_k)$ is the similarity between the actual memory value and the corresponding element for chunk slot k, and is scaled by the mismatch penalty (MP, which was set at 2.0; discussed below). The term ε_i represents transient noise, a random value from a logistic distribution with a mean of zero and variance parameter s of 0.25 (common ACT-R value, e.g., Lebiere, 1999), and introduces stochasticity in retrieval.

The probability of retrieving a particular instance is determined according to the SoftMax equation (i.e., the Boltzmann equation), reflecting the ratio of an instance's activation A_i and the temperature t (which was set to the default value which scales to the noise parameter, $\sqrt{2} * s$):

$$P_i = \frac{e^{A_i/t}}{\sum_j e^{A_j/t}} \quad (2)$$

The model uses ACT-R's *blending* mechanism (Lebiere, 1999, Gonzalez et al., 2003) to generate a classification based on the similarity to past instances. Blending is a memory retrieval mechanism that returns a consensus value across all memories with similar elements, rather than from a specific memory, and is computed by the following equation:

$$\underset{V}{\operatorname{argmin}} \sum_{i} P_i \times \left(1 - Sim(V, V_i)\right)^2 \quad (3)$$

The value V is the one that best satisfies the constraints among actual values V_i in the matching chunks i weighted by their probability of retrieval P_i. Satisficing is defined as minimizing the dissimilarity between the consensus value V and the actual answer V_i contained in chunk i.

In summary, the model matches memories to the current email content and uses blending to generate the classification decision. After generating a classification, the experience (email content plus decision) is saved in declarative memory as a new instance, which affects future decisions. During the training phase, the classification slot is first updated to match the feedback prior to being saved to memory.

While prior research has identified relevant features for detecting phishing emails (Kumaraguru et al., 2009; Singh et al., 2020), and training tools have attempted to teach end-users to identify such features, the current model relies solely on the semantic features of the email to make classifications. At the lowest level, an end-user that has not undergone training to identify expert features would likely rely on the semantics of an email to make a classification. An email that is semantically similar to past known phishing emails would more likely be classified as phishing. Additionally, by relying on only the semantics of the email, the model does not need to identify expert features in a body of text (which is a difficult task to automate by any natural language processing, NLP, standards). In contrast, current NLP techniques are quite efficient at computing the semantic similarity between texts and can therefore feasibly be used to generate the similarities between emails required for blending computations.

A novel feature of the model, therefore, is how similarities are computed between slot values. Typically, similarities between numeric values are computed using a linear function scaled between 0 and 1.0, where 1.0 is a perfect match and 0 is maximally dissimilar. However, for non-numeric information, unless a value is specified for a relation, they are either maximally similar or maximally different. For emails, the content is non-numeric, often several words to paragraphs in length. Because two texts that are semantically similar should have higher similarity values (closer to 1.0) compared to texts that are semantically very dissimilar, it is possible to compute individual similarities between semantic content.

To compute similarities between textual information, we used the University of Maryland Baltimore County's semantic-textual-similarity tool (UMBC; Han et al., 2013). The tool uses a combination of latent semantic analysis (LSA) and WordNet to produce semantic similarity values between two texts. The two input texts can be of any word-length and it produces a value between 0.0 and 1.0, with 1.0 being maximally similar in meaning. For example, the similarity between "happy dog" and "joyful puppy" is 0.65, whereas "happy dog" and "sad feline" is 0.34, and "happy dog" and "hot tea" is 0.0. This technique has proven useful for producing meaningful similarity values between textual content.

Model Results

To generate stable estimates of performance compared to that of humans, the model was run 10 times per participant and given the same sequence of emails presented to the participant. Therefore, in the analyses below, we compare 2980 model runs to 298 humans. Before beginning the task, the model must first be initialized with a set of instances to be able to retrieve a classification. Therefore, the model was initialized with 10 instances that include the email content and ground-truth classification, five of which were phishing emails and five were ham. The initialized instances were sampled from the remaining emails that were not presented during the task.

To examine the model performance compared to that of humans, we computed signal detection measures and plotted the receiver operating characteristic (ROC) curve for each phase and frequency condition of the task. We plotted the mean True Positive Rate (TPR; or Sensitivity) on the y-axis and the False Positive Rate (FPR; or 1-Specificity) on the x-axis. The TPR is equivalent to the hit rate of classifying phishing emails as phishing. The FPR is equivalent to the false-alarm rate of classifying ham emails as phishing. Therefore, in ROC space, points closer to the top left of the graph indicate greater discriminability while points toward the middle indicate less discriminability. Meanwhile, points toward the top right or bottom left indicate greater overall bias toward responding phishing or ham, respectively.

Figure 1 shows the mean ROC curves for the humans (black) compared to the model (gray). As can be seen, the model generates very accurate predictions of human behavior

across phases and frequency conditions. Like humans, the model does not perform perfectly, highlighting the difficulty of the task in discriminating phishing from ham emails. As observed in humans, the frequency of phishing emails observed during training (Phase 2) had a direct impact on discriminability in the post-test phase (Phase 3), such that greater increases in frequency during Phase 2 led to greater increases in TPR, but also FPR, in Phase 3 compared to Phase 1. However, as can be observed, the model is slightly more sensitive to frequency effects than are humans. When the base rate is 25% (10 phishing, 30 ham) the model tends to underpredict human performance at post-test and classifies more of the phishing emails as ham. When the base rate is 50% phishing emails or more, then the model tends to more accurately classify the phishing emails compared to humans. The model demonstrates that a greater frequency of experience with phishing emails leads to more cautious decisions with future emails. This is because the greater number of phishing instances in memory the greater influence they have on retrieval (i.e., a greater probability of retrieving a phishing classification from memory).

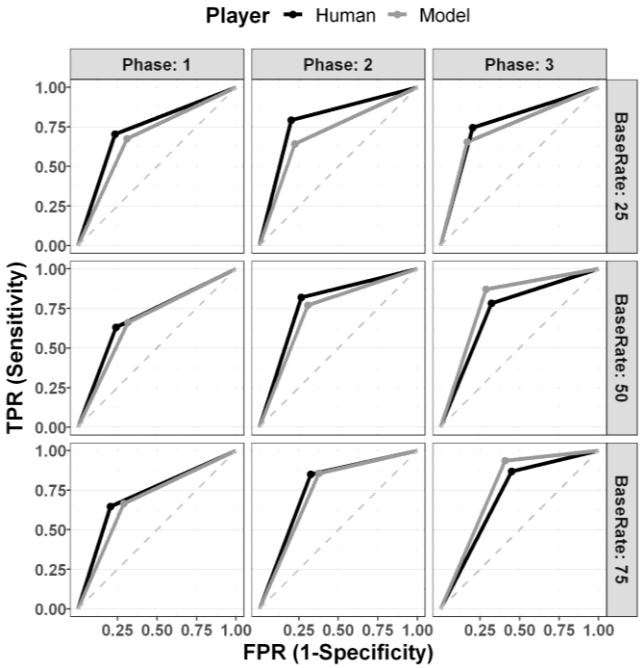

Figure 1: ROC curves of phishing decision accuracy across three phases of the PTT and three frequency conditions, for humans (black) compared to the model (gray).

Discussion

That the model generated highly accurate predictions of human behavior is good news towards developing personalized anti-phishing training interventions. The model is able to rely on experience, through interaction with the environment, and the dynamics of memory to generate a range of behavior. The modifications we made to the original Cranford et al (2019) model helped to provide a better understanding of end-user susceptibility to phishing emails.

For the original model, the important parameter values for activation and blending were left at their default values. These include, decay rate *d*, mismatch penalty *MP*, transient noise *s*, and temperature *t*. For the decay rate, the default value is 0.5. Decay rate is related to forgetting and influences recency effects such that the higher the value the less of an impact older instances will have in retrieval, and thus more recent instances will have a greater impact. At the default value, the model tended classify emails as ham due to the greater frequency of ham emails during the pre-test phase. In the current model, we set this parameter to 0. This allowed all instances to play a more equal role in retrieval and reduced excessive recency effects. Studies have shown that the default decay rate of 0.5 is effective for modeling the typical laboratory task that is short in duration and involves novel stimuli, however, for longer duration tasks this value is less useful at representing retrieval effects (Pavlik & Anderson, 2005). For the present task, reading emails is a task with which humans come to the experiment with vast amounts of prior experience. It is presumable then that these past experiences play a role in retrieval and have decayed to a steady level at experimentation. With such a vast memory base, spreading activation more evenly across new instances by setting decay to zero is a suitable solution for representing this memory phenomenon.

Ideally, we would use a portion of the end-user's actual history of emails to initialize the model, but using examples from the database was a reasonable alternative. Interestingly, the model was initialized with equal numbers of ham and phishing emails, whereas in reality, humans see many more ham emails than phishing. Initializing the model with comparatively more ham emails resulted in too many ham classifications. It could be that phishing emails are inherently more emotionally valent and thus more salient in memory. Or it could be an experimental effect of end-users expecting some of the emails to be phishing. In any case, these results reveal a bias to respond phishing in the task that was accounted for through initialization of instances.

One issue with the original model was that the UMBC semantic similarity tool produces a compressed range of values, which in turn compresses the range of differences between emails and makes it more difficult to discriminate stimuli. During retrieval, the instances are more evenly weighted. To alleviate this constraint, we modified the temperature and mismatch penalty parameters. The noise *s* was left at its default value of 0.25 which provides a reasonable amount of stochasticity in retrieval. Increasing this value resulted in overly varied responses, and reduced discriminability. However, lowering this value did not produce enough stochasticity between model runs. The temperature on the other hand was reduced from a neutral value of 1.0 to the default value of $\sqrt{2} * s$, which equals approximately 0.35 given the current value of *s*. Temperatures of 1.0 reflect an unbiased retrieval given the historical frequency distribution of instances. This means that retrieval is

more evenly distributed across instances. Increasing this value tends toward randomness, making discrimination more difficult. Therefore, lowering the temperature by reverting to the default ACT-R value resulted in greater discriminability where more weight is given to instances with higher activation values. This in turn rewards those instances that are more semantically similar to the current instance. Finally, we increased the mismatch penalty *MP*, from 1.0 to 2.0. The mismatch penalty directly influences the model's discriminability because it scales the dissimilarity between instances. Therefore, increasing this value enhanced the differences between different emails while simultaneously strengthening the similarities between similar emails, effectively decompressing the range of similarities produced by the UMBC semantic-similarity tool. The result was an increase in overall discriminability of the model.

The current model predicts human performance well, but even still, there is room for improvement. We did not perform any detailed parameterization of the model, but instead settled on reasonable and justifiable values through strategic exploration. Therefore, the model has potential to be further refined. Additionally, relying on the semantic similarity of features of an email, generated through NLP techniques, instead of attempting to extract the presence of features within the email text, allowed us to create a model that can more easily generalize to novel environments (i.e., with different emails). Relying on the semantic content means we do not have to preprocess new emails, manually or through automated means, to identify relevant features. To test the generality of the model, we ran the model through another task that used a different database of emails, the Phishing Email Suspicion Test (Hakim et al., 2019).

Modeling the Phishing Email Suspicion Test

Hakim et al. (2019) used the PEST task to assess the relationship between real and simulated phishing and ham emails and to examine the efficacy of using the simulated phishing emails for anti-phishing training against real-world phishing attempts. In the PEST task, 97 participants rated a total of 160 emails each on a Likert-type suspiciousness scale from 1 "Definitely Safe" to 4 "Definitely Suspicious". Participants were presented 40 of each type of email: real-ham, simulated-ham, real-phishing, and simulated-phishing. The emails were presented in random order and selected randomly from the database of emails.

The IBL model described above was tasked to perform the PEST. Because the PEST included four types of emails, the model was initialized with a total of 20 emails (five of each type). However, to model the individual differences observed in the human PEST data, we introduced stochasticity in the initialization. That is, the model was initialized with three to six ham emails of each type, randomly selected from a uniform distribution, and the remaining were phishing. This also introduced varied initial biases between model runs, where some runs were initially biased toward ham and other runs more biased toward phishing, thus resembling a human population, but with a skew toward phishing. However, as will be discussed in more detail below, to produce the following results, each run was initialized with an extra 2 simulated-phishing emails. This is consistent with the finding of Hakim et al. (2019) that participants displayed a bias to respond phishing, and with the phishing bias observed in the PTT model. Increasing the number of phishing emails was required to drive such a bias in the PEST model.

Since the PEST database of emails includes only 40 examples of real-ham emails, to ensure no initialized real-ham emails were presented during the test, we reduced the number of emails of each type presented during testing from 40 to 30. Therefore, the model experiences 120 total emails per run, still allowing for ample observations. For the PEST, emails did not show the underlying link URL if hovered over with a mouse, so the URL feature was removed from the instance representation and only the link text was compared in retrievals. Instead of generating a classification, the model takes the semantic features of the email as input and generates, via blending, a rating score between 1 and 4. The retrieved value is a continuous value that is rounded to the nearest whole number to provide a discrete rating. The blended rating value is replaced with the discrete rating value before saving the instance to declarative memory at the end of each trial. Like humans, the model does not receive feedback regarding the accuracy of its decisions.

Model Results

To generate stable estimates of performance, the model ran through the task 200 times, with initialized and tested emails selected randomly for each run. Therefore, in the analyses below, we compare 200 model runs to 97 humans.

Because the PEST requires a rating response as opposed to a classification response, to analyze the model performance compared to humans, we examined the mean suspicion scores per email type as well as the subject-level correlation between ratings for simulated and real emails, separately for phishing and ham emails. These combined measures provide accounts for the mean as well as the full range of human behavior.

Figure 2 shows a boxplot of the mean suspicion score per email type. The results align very well with the human data, closely accounting for the mean behavior as well as the variance between individuals. The real-ham emails were rated the lowest at approximately 2, while the simulated- and real- phishing emails were rated highest at almost 3. Meanwhile, the simulated-ham emails were rated at near 2.5.

Figure 3 shows the correlation between simulated and real emails for ham and phishing emails separately. These results highlight the model's ability to account for both the within- and between-subject variances in performance. As will be discussed further, a key contributor to the model's performance is the randomization of initial instances.

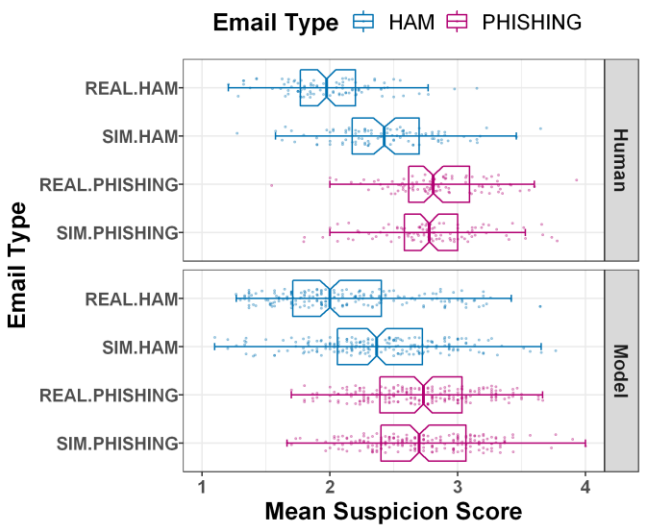

Figure 2: Boxplot of mean suspicion scores for each type of email in the PEST for humans compared to the model.

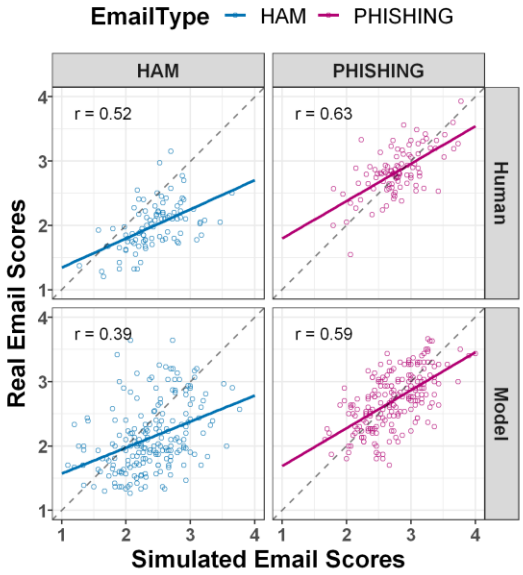

Figure 3: Scatterplots showing the correlation between real emails and simulated emails for ham emails and phishing emails separately, comparing humans to the model.

Discussion

The IBL phishing susceptibility model was able to successfully generalize to other environmental conditions with a different pool of participants performing a slightly different task with different stimuli. To produce the level of accuracy in predicting human behavior in the PEST, the model required stochasticity in initialization and additional initialized phishing instances. The result was an increase in correlations between real and simulated emails and an overall bias toward rating emails more suspicious. In fact, using a static initialization considerably reduced the observed correlations in Figure 3.

In exploring an appropriate initialization for the model, the results revealed a relationship between simulated-ham and -phishing emails. For example, increases in simulated-phishing emails had a positive impact on simulated-ham emails. These results suggest there is large semantic overlap between simulated emails, which is consistent with how the simulated-phishing and -ham emails were constructed. The simulated-ham emails were modified versions of simulated-phishing emails made to seem less suspicious. The model picks up on this semantic overlap which results in simulated-ham emails having a higher match to simulated-phishing emails, producing inflated ratings for simulated-ham emails.

Conclusion

Our IBL model highlights the role of experiential learning for end-user phishing detection decisions. The major influences in generating accurate predictions of human susceptibility to phishing emails were initialization of instances and similarity between email features. A phishing bias was accounted for by adding disproportionally more phishing emails than ham emails compared to real-world frequencies. Adding stochasticity in initialization accounted for individual differences in behavior. Humans have distinct experiences that influence decisions and capturing this background knowledge is essential to building models that not only predict a range of human behavior, but also that can predict a specific individual's behavior. Because the model is expected to generate better predictions of an individual the more the model's memory aligns with the human's, model-tracing techniques should prove useful in developing personalized anti-phishing training interventions (Anderson et al., 1995). Future research is aimed at further exploring the effects of initialization, with an emphasis on generality and in exploring ways to decompress the range of semantic similarities or even representing alternative features.

While the current model uses only the semantics of the email to make decisions, current training methods teach end-users to identify so-called "expert" features (e.g., a request for personal information; Singh et al., 2020). Using only semantic features of an email produces human-like, albeit fairly poor discriminability in the experimental tasks. A goal for future research is developing a model that can learn to identify expert features so that we can use the model to help train end-users to detect such features. For now, the current model proved a successful first step toward personalized anti-phishing training.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant Number 2026148, the Carnegie Mellon University CyLab Security and Privacy Institute, and the Army Research Office under MURI Grant Number W911NF-17-1-0370.

References

Anderson, J. R., Corbett, A. T., Koedinger, K., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of Learning Sciences, 4*, 167–207.

Anderson, J. R., & Lebiere, C. (1998). *The Atomic Components of Thought*. Mahwah, NJ: Erlbaum.

Anti-Phishing Working Group (2021). *Phishing Activity Trends Report: 4th Quarter 2020.* Retrieved from https://docs.apwg.org/reports/apwg_trends_report_q4_2020.pdf. APWG.

Canfield, C. I., Fischhoff, B., & Davis, A. (2016). Quantifying phishing susceptibility for detection and behavior decisions. *Human factors, 58*(8), 1158-1172.

Cranford, E. A., Lebiere, C., Rajivan, P., Aggarwal, P., & Gonzalez, C. (2019). Modeling cognitive dynamics in end-user response to phishing emails. In *Proceedings of the 17th Annual Meeting of the International Conference on Cognitive Modelling*. Montreal, CA.

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological review*, *118* (4), 523.

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance based learning in dynamic decision making. *Cognitive Science, 27*(4), 591-635.

Hakim, Z.M., Ebner, N.C., Oliveira, D.S. *et al.* (2020). The Phishing Email Suspicion Test (PEST) a lab-based task for evaluating the cognitive mechanisms of phishing detection. *Behavioral Research Methods*.

Han, L., Kashyap, A. L., Finin, T., Mayfield, J., & Weese, J. (2013). UMBC_EBIQUITY-CORE: Semantic Textual Similarity Systems. In *Proceedings of the 2nd JCLCS* (pp. 44-52). Atlanta, GA.

Jagatic, T. N., Johnson, N. A., Jakobsson, M., & Menczer, F. (2007). Social phishing. *Communications of the ACM, 50*(10), 94–100

Kumaraguru, P., Cranshaw, J., Acquisti, A., Cranor, L., Hong, J., Blair, M. A., & Pham, T. (2009). School of phish: a real-world evaluation of anti-phishing training. In *Proceedings of the 5th Symposium on Usable Privacy and Security* (p. 3). Mountain View, CA.

Kumaraguru, P., Rhee, Y., Sheng, S., Hasan, S., Acquisti, A., Cranor, L. F., & Hong, J. (2007). Getting users to pay attention to anti-phishing education: evaluation of retention and transfer. In *Proceedings of the anti-phishing working groups 2nd annual ecrime researchers summit* (pp. 70-81).

Lebiere, C. (1999). A blending process for aggregate retrievals. In *Proceedings of the 6th ACT-R Workshop*. George Mason University, Fairfax, Va.

Lebiere, C., Pirolli, P., Thomson, R., Paik, J., Rutledge-Taylor, M., Staszewski, J., & Anderson, J. R. (2013). A Functional Model of Sensemaking in a Neurocognitive Architecture. *Computational Intelligence and Neuroscience.*

Lin, T., Capecci, D. E., Ellis, D. M., Rocha, H. A., Dommaraju, S., Oliveira, D. S., & Ebner, N. C. (2019). Susceptibility to Spear-Phishing Emails: Effects of Internet User Demographics and Email Content. *ACM Trans. Comput.-Hum. Interact.* 26, 5, Article 32 (July 2019), 28 pages.

Marchal, S., François, J., State, R., & Engel, T. (2014). Phishstorm: Detecting phishing with streaming analytics. *IEEE Transactions on Network and Service Management*, *11*(4), 458-471.

Oliveira, D., Rocha, H., Yang, H., Ellis, D., Dommaraju, S., Muradoglu, M., Weir, D., Soliman, A., Lin, T., & Ebner, N. (2017). Dissecting Spear Phishing Emails for Older vs Young Adults: On the Interplay of Weapons of Influence and Life Domains in Predicting Susceptibility to Phishing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (*CHI '17*). Association for Computing Machinery, New York, NY, USA, 6412–6424.

Pavlik Jr, P. I., & Anderson, J. R. (2005) Practice and forgetting effects on vocabulary memory: an activation-based model of the spacing effect. *Cognitive Science*, 29(4), 559–586.

Peng, T., Harris, I., & Sawa, Y. (2018). Detecting phishing attacks using natural language processing and machine learning. In *Proceedings of the IEEE 12th international conference on semantic computing* (pp. 300–301).

Prakash, P., Kumar, M., Kompella, R. R., & Gupta, M. (2010). Phishnet: predictive blacklisting to detect phishing attacks. In *2010 Proceedings IEEE INFOCOM* (pp. 1-5). San Diego, CA.

Singh, K., Aggarwal, P., Rajivan, P., & Gonzalez C. (2019). Training to detect phishing emails: Effect of the frequency of experienced phishing emails. In *Proceeding of the 63rd International Annual Meeting of the HFES.* Seattle, WA.

Singh, K., Aggarwal, P., Rajivan, P., & Gonzalez, C. (2020). What makes phishing emails hard for humans to detect? In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 64*(1). Chicago, IL.

Vishwanath, A., Harrison, B., & Ng, Y. J. (2018). Suspicion, Cognition, and Automaticity Model of Phishing Susceptibility. *Comm. Research, 45*(8), 1146–1166.