Modeling Cognitive Dynamics in End-User Response to Phishing Emails

Edward A. Cranford (cranford@cmu.edu) and Christian Lebiere (cl@cmu.edu)
Department of Psychology, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213 USA

Prashanth Rajivan (prajivan@uw.edu)
Department of Industrial & Systems Engineering, University of Washington
3900 E Stevens Way NE, Seattle, WA 98195 USA

Palvi Aggarwal (palvia@andrew.cmu.edu) and Cleotilde Gonzalez (coty@cmu.edu)
Dynamic Decision Making Laboratory, Social and Decision Sciences Department, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213 USA

Abstract

Phishing attacks are a significant threat to cybersecurity, while current defense methods do not adequately address the human factor of this threat: the role of experiences and cognitive biases. To better understand human susceptibilities to phishing attacks, we developed an Instance-Based Learning (IBL) model for predicting end-user's behavior in a phishing email detection task. We present a phishing scenario that demonstrates that typically safe end-users can fall victims to phishing attacks in certain circumstances, and these situations are the result of cognitive mechanisms such as frequency and recency and similarities between memory events. We demonstrate the ability of an IBL model to predict human performance in a laboratory phishing detection task. While the results indicate that phishing detection was difficult for the model, it roughly reflects in the data the difficulty humans had. Future research is aimed at enhancing the IBL model to better predict end-user phishing detection, and to explore the ways in which this model can be used as a training tool and online aid for end-user detection of phishing attacks.

Keywords: phishing; cybersecurity; decision making; instance-based learning; cognitive model; ACT-R

Introduction

All it takes is one click in response to a phishing email to compromise the security posture of an entire organization, and as such phishing attacks pose the biggest threat for cybersecurity (Wombat Security report, 2018). Phishing aims to persuade end-users to share sensitive information using social engineering and psychological techniques (Jagatic et al., 2007). While phishing attacks exploit human weaknesses, defenders typically employ technological solutions to defend against them, such as machine learning filtering of phishing emails, email authentication tools, URL filtration, and blacklisting phishing URLs (Prakash et al., 2010; Marchal et al., 2014; Peng, Harris, & Sawa, 2018). Current methods of defense against phishing attacks are insufficient because they don't consider human cognitive biases and experience. Since the success of phishing attacks rely on the exploitation of end-user's cognitive and psychological weaknesses, it becomes essential to understand the detection capabilities, decision making, and cognitive biases of end users who respond to phishing emails (Canfield, Fischhoff, & Davis, 2016).

Considerable research has been devoted to investigating how to best train end-users to detect phishing emails (Kumaraguru et al. 2009, Jensen et al., 2017), yet even trained end-users can still fall victim to phishing attacks. Recent research examining the interaction between attackers and end-users revealed various strategies that attackers use to design phishing campaigns and their success on end-user's detection of phishing emails (Rajivan & Gonzalez, 2018; Curtis et al., 2018, Singh et al., 2019). In the current research, based on psychological theories of decisions from experience, and the insights of these recent phishing studies, we propose a cognitive model of end-user phishing email detection. Our insights suggest that phishing emails detection is influenced by the end-user's prior history of emails, their recent experiences, and their innate and learned cognitive biases.

In what follows, we first describe an example phishing scenario that reveals the process by which an end-user might fall trap to an attacker's social engineering strategies. We then formalize a cognitive model of end-user email classification, built in the ACT-R cognitive architecture (Anderson & Lebiere, 1998), using Instance-based Learning Theory (IBLT; Gonzalez, Lerch, & Lebiere, 2003). Using the data set from Rajivan and Gonzalez (2018), we demonstrate that cognitive models of end-user detection of phishing attacks can be useful for understanding how and when humans are most vulnerable to attacks, providing insights on how to best train people to detect phishing emails, and could potentially serve as a powerful decision support tool to prevent phishing attacks.

A Cognitive Model of Phishing Email Detection

In the example phishing scenario, depicted in Figure 1, Alice is a representative persona for a class of members of a fictional organization. The cyber-security division is assessing vulnerabilities of phishing attacks and sends Alice a number of emails, some of which are phishing emails. Her task is to decide whether to click a link within an email.

Alice represents a particularly savvy end-user, who usually recognizes malicious emails, and does not click on embedded links. In this scenario, Alice starts with a prior history of not clicking links from unknown senders (i.e., senders that she has not previously interacted with and whom she does not

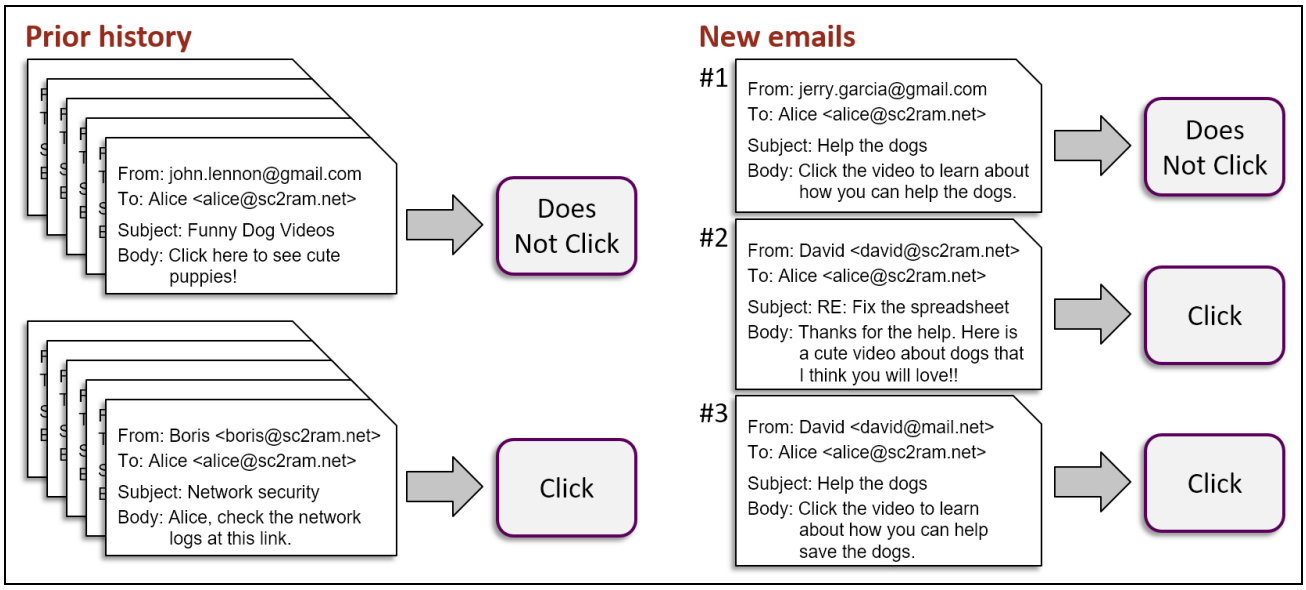

Figure 1: The Alice scenario: an example phishing attack scenario.

recognize). She has a predisposition to click on links from trusted senders (e.g., coworkers and friends), particularly about topics that interest her (e.g., one of Alice's interests is in dogs). She is then presented with three new emails, one at a time. The first is from an unknown sender about dogs, for which she does not click the link. The second is from a trusted coworker that mentions dogs, for which she clicks the link. In the third email, after observing and/or inferring Alice's clicking behavior, the attacker spoofs the sender's source address, pretending to be a colleague of Alice, and baits her with a topic and domain name related to dogs. Alice clicks the link and self-compromises her system.

Alice's behavior can be described as emerging from the interaction between her learned behavior/tendencies and changes to the environment. The cognitive model, described next, captures underlying cognitive mechanisms such as priming, transfer, and recency bias that reflect the statistics and dynamics of the environment and give rise to Alice's behavior. As shown in Figure 2, Alice's prior history of emails may cluster on dimensions of email topic (work, dogs) and sender (known, unknown). Emails about dogs from unknown senders cluster together and embedded links are typically not clicked. Whereas, emails about work topics from known coworkers cluster together, and embedded links are typically clicked. The first email is similar to past emails for which she did not click on embedded links, and so she doesn't. The second email is from trusted coworkers, but mentions dogs, yet is more similar to past emails for which she clicked on links, and so she does. This expands the cluster of emails for which she previously clicked. Alice would typically not click on the link in the third email, because it is more similar to past emails for which she did not click embedded links. However, it is more similar to the recent second email, and so is pulled toward the cluster of emails for which she clicked links. Alice's normal behavior has changed as a result of her interactions with the environment over time.

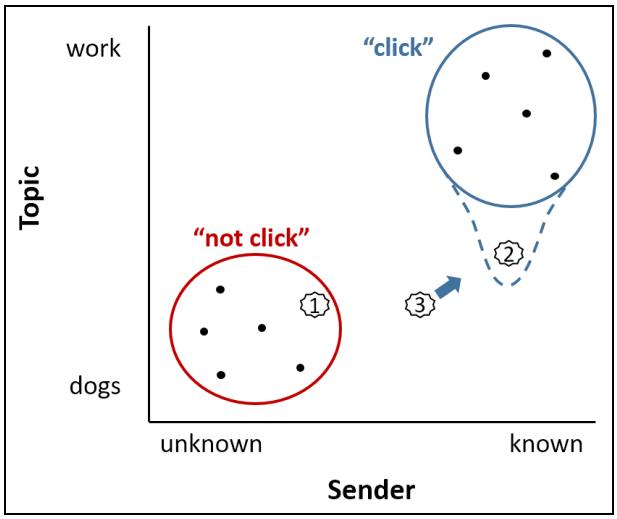

Figure 2: Representation of Alice's behavior.

An IBL Model of Phishing Detection

According to IBLT (Gonzalez et al., 2003), decisions are made by generalizing across past experiences, or instances, that are similar to the current situation. Typically, instances are encoded as chunks in declarative memory that represent the features of the decision: the context in which a decision is made, the action taken, and the outcome of that decision. For emails, there is usually a dissociation between the actions taken and feedback regarding whether the email was ultimately malicious. Therefore, for this task, only the context and the action are represented within each instance, but not the outcome. The context elements of an email include the sender's email address, the subject line, the body of the email, and the link. The action slot includes the action taken (either click or not click the link). Initial past instances include those represented in Figure 1 under *Prior History*: five emails from

unknown senders about various topics, including puppies, for which Alice did *not* click on the embedded links, and five emails from trusted coworkers, about work-related topics, for which Alice clicked on the embedded links.

An IBL cognitive model was constructed in the ACT-R cognitive architecture (Anderson & Lebiere, 1998). For each new incoming email (see Figure 1, "New Emails"), the model takes as input the context of the email and generates an action by retrieving similar past instances. 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 context. The activation A_i of a chunk i is determined by the following equation:

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

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 which is set to the default ACT-R value of 0.5. 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 context element for chunk slot k, and is scaled by the mismatch penalty (MP) which was set to the default value of 1.0. 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), to introduce 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 1.0):

$$P_i = \frac{e^{A_i/t}}{\sum_j e^{A_j/t}}$$

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

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

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 context and uses blending to generate the action. After generating an action, the experience (context plus action) is saved in declarative memory as a new instance, which affects future decisions.

An important feature of the model is how similarities are computed between slot values. Typically, similarities between numeric values are computed on a linear function scaled between 0 and -1.0, where 0 is a perfect match and -1.0 is maximally dissimilar. However, for non-numeric information, unless a value is specified for relation, they are either maximally similar or maximally different. For emails, the context is non-numeric, often several words to paragraphs in length. It is sensible then that two texts that are semantically similar should have higher similarity values (closer to 0) compared to texts that are semantically very dissimilar.

In order to compute similarities between slot contents involving textual information, we used the University of Maryland Baltimore County's semantic-textual-similarity tool (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. We subtract one from this value to produce a dissimilarity value for use in blending. This technique has proven to be a useful methodology for producing meaningful similarity values for textual content.

Demonstration of the IBL Model Behavior

Figure 3 shows the model behavior during a typical run through the Alice scenario. The first column shows the new incoming emails. The second column shows Alice's prior history of emails stored in memory: the top stack shows emails for which Alice previously did not click on the embedded link, while the bottom stack shows emails for which she did click. For each new email, the model retrieves a decision based on its similarity to prior emails. The darker the email, the less recent it was experienced and encoded in memory. Darker, fuller arrows indicate greater activation strength (purple) or decision weighting (orange). The third column shows the blending values (i.e., the relative weighting given to each option based on activation and similarities) next to the two possible decisions (Click or Not-Click). The decision made is that with the greater blending value.

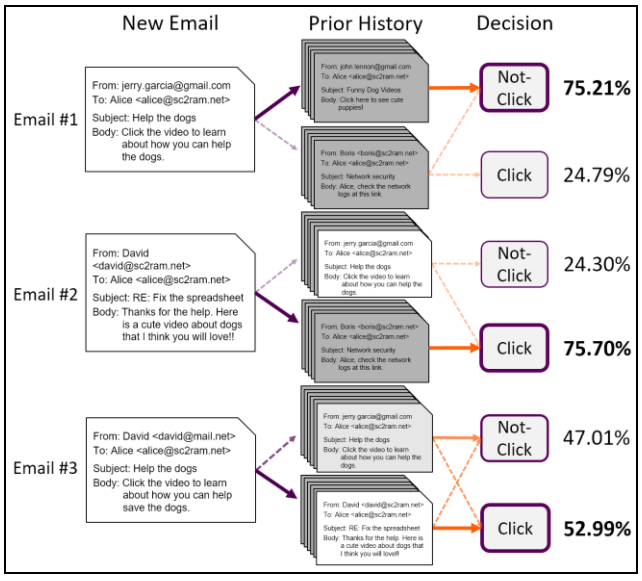

Figure 3: Example model behavior in the Alice scenario.

For example, for Email #1, its context is more similar to past emails from unknown senders than to those from trusted coworkers. The blending mechanism produces a distance metric from each possible decision, and since there are only two possible decisions, blending values can be translated directly into the weighted probability of making each decision. Therefore, for this particular run, the model decides to not click the link with a weighting of 75.21%. For Email #2, its context is more similar to emails from trusted coworkers, and the model decides to attack with a weighting of 75.7%. For the critical Email #3, although the contents are typically more similar to unknown senders, it also shares similarity with the most recent email from a trusted coworker giving more weight to the decision to click on the link (e.g., recency bias to click on a link about dogs from a David, because of the similarity to the contents of Email #2). On this run, the combination of activation strength and similarity across past instances result in a decision to click on the link with a weighting of 52.99%.

This example shows how under certain circumstances a relatively safe user could sometimes get caught performing an unsafe act. To generate stable predictions of human behavior, the model was run 1000 times to highlight its activation dynamics. For Email #1, the model decided to click the link on less than 1% of the 1000 runs, with a mean weighting in favor of not clicking of 66.3%. For Email #2, the model clicked the link on 98.8% of runs, with a mean weighing in favor of not clicking of 66.2%. For Email #3, the model clicked the link on 56.6% of runs, with a mean weighing in favor of clicking of 50.8%. Of course, that action itself will make further dangerous actions more likely.

The IBL model of the Alice phishing scenario shows how a user's response to phishing emails may be highly constrained by cognitive mechanisms, especially activation in declarative memory, which reflect the statistics and dynamics of the environment in the user's memory. Alice's behavior is a result of manipulating that environment in a way that can change well established behaviors. As demonstrated, it only takes a short history of human behavior, and their interests, to personalize a model to an individual user and make predictions about whether the user might perform an unsafe act when encountering a malicious email.

Validation of the IBL Model Against Humans

To assess performance of the IBL model described above, it was adapted to predict human behavior in a laboratory experiment, reported in Rajivan and Gonzalez (2018). Their data set includes 340 participants as end-users in an email management task. Participants were presented with 20 emails, one at a time; 10 were benign emails and 10 were phishing emails, randomly distributed. Their task was to assist a fictional office manager by examine each of her incoming emails and decide how to respond: 1) respond immediately; 2) flag the email for follow up; 3) leave the email in the inbox; 4) delete the email; or 5) delete the email and block the sender. An email rated as 1 can be viewed as more benign and important, while an email rated as 5 is more malicious.

For this task, the chunk definitions of the model were modified to represent the information available to participants. For these emails, there was no sender information available, but links were represented both as the HTML link as well as the observable text in the email. Therefore, the context slots include the subject, body, link, and link text. The decisions were recoded to be analogous to the conceptual model, with ratings of 1 and 2 recoded as "respond" (i.e., the equivalent of clicking a link) and ratings of 3 through 5 recoded as "do not respond" (i.e., the equivalent of not clicking a link). Therefore, for the model, the possible decisions are *respond* or *not-respond*. All parameters were left the same as for the conceptual model.

Results

The model was run 10 times for each participant and was presented the same stimuli experienced by the human. The first 10 emails experienced served as training instances for the model and were encoded as an initial declarative memory. The model then made a decision for each of the next 10 emails, and its predictive accuracy was evaluated.

The model performed better than chance (50%), accurately predicting the human's decision on 58.6% of benign emails and 63.4% of phishing emails, on average. The model was more accurate on phishing emails than benign emails, $F(1,9) = 10.12$, $p = 0.001$. There were no differences across trials and the interaction was not significant, both p's > 0.43.

The confusion matrices presented in Figure 4, show the percentage of trials in which the model and human agreed in their decisions to respond to the email or not, for phishing emails (top) and benign emails (bottom). D-prime for phishing emails is 0.60, while it is 0.43 for benign emails. Figure 4 also shows the phishing detection accuracy of humans and the model. For both phishing and benign emails, the model and humans decided to respond to ~40% of emails or more. As a result, the model more accurately predicts human decisions to *not* respond to an email than to respond.

Like humans, the model responded to a large proportion of phishing emails (39.7% and 39.0% respectively). Although, while humans responded to more benign emails (47.9%), the model responded to only 39.5% of benign emails – almost the same rate as phishing emails, indicating that distinguishing between ham and phishing emails was difficult.

Discussion

Humans were less cautious in the email management task than they might normally be in real-world circumstances, and the IBL model reflected this behavior, and responded to many phishing emails. Overall, the model was better than chance at predicting human performance, but the task proved difficult for both the humans and the model without rewards or feedback to aid learning. The model was trained on the first 10 trials of human data, and therefore reflects the overall tendencies to not respond. However, while the model is similarly as biased as humans to not respond to emails, it has a slightly more difficult time distinguishing a benign email from a phishing email than humans.

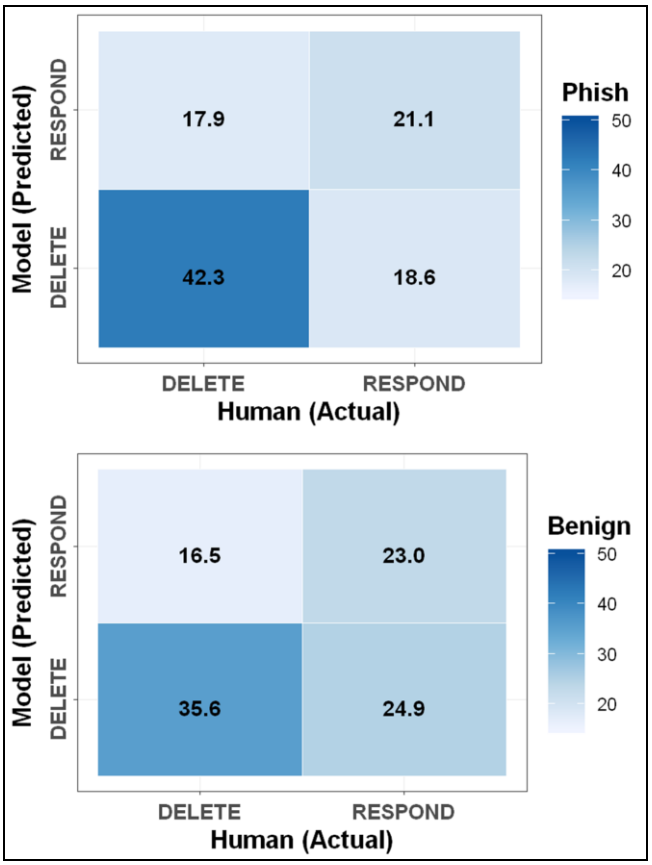

Figure 4: Confusion matrices comparing the model predictions of human decisions (in percentages) for phishing emails (top) and benign emails (bottom).

The benign emails in Rajivan and Gonzalez (2018) were ham emails that came from businesses and senders, and were about topics and accounts, that the end user could not know were relevant to their fictional office manager. Without context, the benign emails look and sound very similar to the phishing emails, making the detection of phishing emails difficult. In fact, when looking into the UMBC similarities within and between benign and phishing emails, the values are very close to each other. The mean similarities between benign emails are the highest, but still relatively low, at only 0.43. Meanwhile, the phishing emails are as dissimilar to each other (0.36) as they are to benign emails (0.39). The model accurately captures overall human tendencies, but has more difficulty than humans in classifying a benign email as safe.

Limitations and Future Directions

There is clear room for improvement for the IBL model. It is limited by its representation of the relevant features for detecting phishing emails. Research in human susceptibility to phishing scams has revealed important cues and indicators of phishing emails that end-users should be trained to detect (Vishwanath, Harrison, & Ng, 2018). While the sender, subject line, URL, and the email body are all important features to use for detection, representing only the semantic content limits the model's ability to discriminate. Some features that could be extracted from the email to enhance the representation include grammar and spelling ratings, emotional tone of the email, and sentiment. Future research is aimed at exploring and expanding the features that represent the context of an email. It is also unclear at this point what features the model relies on most to make decisions or if any do not affect decisions. The features need to be accurately represented in the context to accurately reflect the statistical dynamics of the environment. Representing user interests, as well as background knowledge of known senders, are additional features that would greatly improve the model's ability to predict a particular individual's behavior.

On a related note, while the UMBC semantic similarity tool proved useful, many of the similarity values between emails are in the range of 0.33-0.66. Adjusting these values so they fill the full range of 0.0-1.0 could help to increase the dissimilarity between the benign and phishing emails while increasing the similarity within email types. Additionally, the similarities are computed between entire email bodies. These bodies could be parsed into separate phrases to uncover more fine-grained features. Future research is aimed at exploring these possibilities.

Improving the cognitive model of phishing detection is an important goal for gaining a better understanding of end-user susceptibility to phishing emails. Additionally, there is a wide array of possible applications in cybersecurity, including using cognitive models to help train end users to detect phishing attacks. A cognitive model that can track a user's experience helps reveal instances when a user may be more susceptible to a phishing scam. The model can make the user aware of such instances to improve their detection. Predicting individual end user behavior is a challenging task but could be extremely helpful in aiding end users in online detection.

After improving the cognitive model, the model can be scaled up to larger applications. For example, cognitive models could also be used to estimate the risk of new phishing samples, or as part of a larger simulation testbed for cyber defense exercises, or to test tools. For applications such as these, scalability becomes an issue for computing semantic similarities. Tools like UMBC's similarity tool typically look up information from very large databases. If you only need to compute a few values per iteration, then computation costs are minimal. However, computation time increases exponentially as the number of instances in the model's declarative memory increases. One technique that proved useful for us was to build a hash-table that stores similarity values between two phrases, thereby eliminating the need to re-compute values for distinct pairs of phrases. If the corpus of emails is known, then these values can be computed prior to running the model. Otherwise, the model would only be able to reuse values after the first experience. Another approach is to use vector embeddings, then compute similarities as distances between vectors.

In the Rajivan and Gonzalez (2018) study, participants saw a large proportion of phishing emails compared to benign emails (50% precisely). Using the same dataset, Singh et al.

(2019) conducted another study to investigate how the frequency of experiencing phishing emails during training affected detection in a later testing phase. Participants completed three phases in a phishing detection task: pre-training, training, and post-training, where participants were trained on different frequencies of phishing emails (25%, 50%, or 75%) and tested before and after training with 20% frequency of phishing emails. The results showed that participants that saw a larger proportion of phishing emails during training had higher hit rates but also higher false alarm rates in detecting phishing emails. This, in addition to the similarity between the benign ham emails and the phishing emails, can explain the bias to not respond to emails in the Rajivan and Gonzalez task. In the future, we will adapt the cognitive model to the task performed in Singh et al. to test other predictions of the IBL model, given that frequency of instances is one of the driving cognitive factors that influence decision making. A similar line of research will explore the model's ability to predict end-user behavior in situations where the statistics of the environment are more similar to that in the real world (e.g., where a very small proportion of emails are phishing emails).

Conclusions

In this paper we demonstrated that a cognitive model of end-user detection of phishing emails can be useful for understanding human susceptibility to phishing attacks. As the Alice scenario showed, normally safe end-users can get caught performing unsafe actions under the right set of circumstances. Human decisions are constrained by cognitive mechanisms (e.g., memory, spreading activation, and pattern matching) that reflect the statistics and dynamics of the environment. By manipulating that environment, new patterns can arise that change well-established user behavior.

The IBL model developed here is a first attempt to model phishing detection using ACT-R, and captures the cognitive mechanisms and biases that could give rise to unsafe actions. It is also a first step toward developing a cognitive model that predicts human performance based on the similarity of emails confronted. According to IBLT (Gonzalez et al., 2003), decisions are based on the similarity of the current email to past emails for which the user clicked links, the recency of those past emails, and the frequency of phishing emails in comparison to benign emails. The model performed similarly to the actions taken by humans, neither the model nor humans were highly accurate in classifying phishing emails. The nature of the task made classification difficult for both. Future research will investigate the various cognitive aspects that influence classification decisions, and improve the context representation in the model to reflect the relevant features for phishing detection. A cognitive model that is highly accurate at predicting end-user susceptibility to phishing attacks can greatly enhance current cybersecurity practice.

Acknowledgments

This research was sponsored by the Army Research Office and accomplished under Grant Number W911NF-17-1-0370.

References

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

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

Curtis, S. R., Rajivan, P., Jones, D. N., & Gonzalez, C. (2018). Phishing attempts among the dark triad: Patterns of attack and vulnerability. *Computers in Human Behavior, 87*(2018), 174-182.

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

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.

Jensen, M. L., Dinger, M., Wright, R. T., & Thatcher, J. B. (2017). Training to mitigate phishing attacks using mindfulness techniques. *Journal of Management Information Systems, 34*(2), 597–626.

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.

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

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.

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.

Rajivan, P., & Gonzalez, C. (2018). Creative Persuasion: A Study on Adversarial Behaviors and Strategies in Phishing Attacks. *Frontiers in Psychology, 9*(135), 1-14.

Wombat Security (2018). *Beyond the Phish Report*. Retrieved from https://www.wombatsecurity.com/beyond-the-phish-2018. Proofpoint Inc.

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