Carnegie Mellon University

Toward Personalized Adaptive Anti-Phishing Training and Automated Assistants

Presented by
Edward A. Cranford[1]

Christian Lebiere[1], Kuldeep Singh[2],
Palvi Aggarwal[2], & Cleotilde Gonzalez[1]
[1]Carnegie Mellon University
[2]University of Texas El Paso

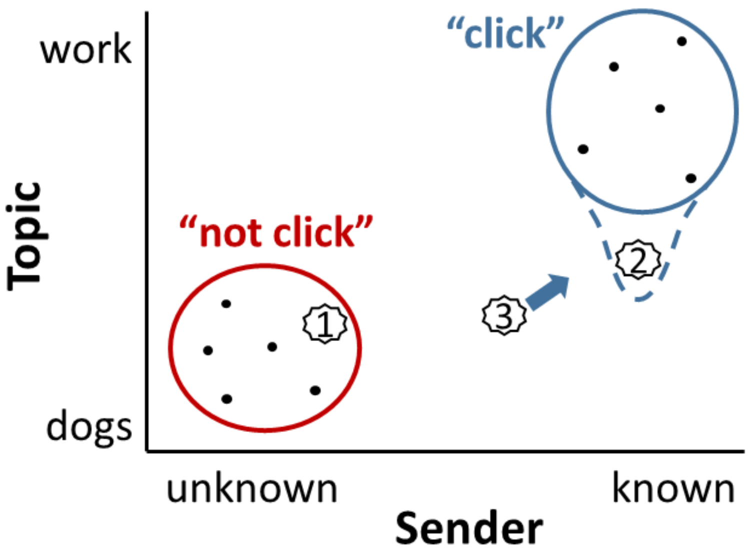
CyLab Carnegie Mellon University Security and Privacy Institute

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.

Computational Cybersecurity in Compromised Environments
2021 Fall Workshop | October 27-28 | Virtual

Using cognitive models to drive personalized, adaptive anti-phishing training systems

- Cognitive models are scalable alternatives to human trainers that can be personalized to an individual to assist them when they deviate from safe behavior
  - e.g., the end-user, the frontline of cybersecurity
- Traditional anti-phishing training is often non-personalized and does not typically account for human experiential learning
  - Personalized training requires accurate models and predictions of individual susceptibility to phishing emails
- We propose that phishing classification decisions are similar to other kinds of decisions from experience
  - Instance-Based Learning (IBL) Theory[1] used to build cognitive models of classification decisions of phishing emails



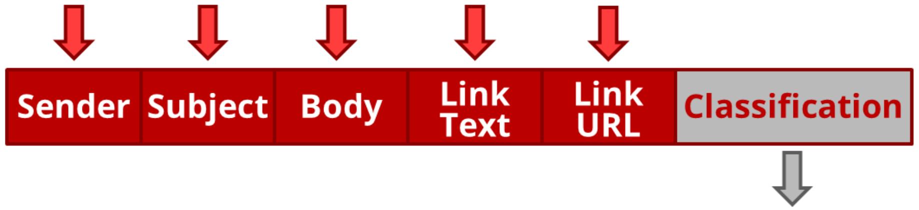
[1]Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance based learning in dynamic decision making. *Cognitive Science*, *27*(4), 591-635

An IBL model of end-user susceptibility to phishing emails

- IBL model built in the ACT-R cognitive architecture
  - Decisions made by retrieving a classification from memory based on the similarity of features of the current email to features of past emails
    - Process generalizes across past experiences
      - (i.e., *blending*[2])
    - Influenced by matching and retrieval mechanisms
      - Similarity of current instance to past instances
      - Recency of past instances
      - Frequency of past instances
  - Similarities based on the semantic similarity between email features
    - Uses NLP technique to automate process
    - UMBC Semantic Similarity Tool[3]

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

$$P_i = \frac{e^{A_i/t}}{\sum_j e^{A_j/t}} \qquad \underset{V}{\operatorname{argmin}} \sum_i P_i \times \left(1 - Sim(V, V_i)\right)^2$$



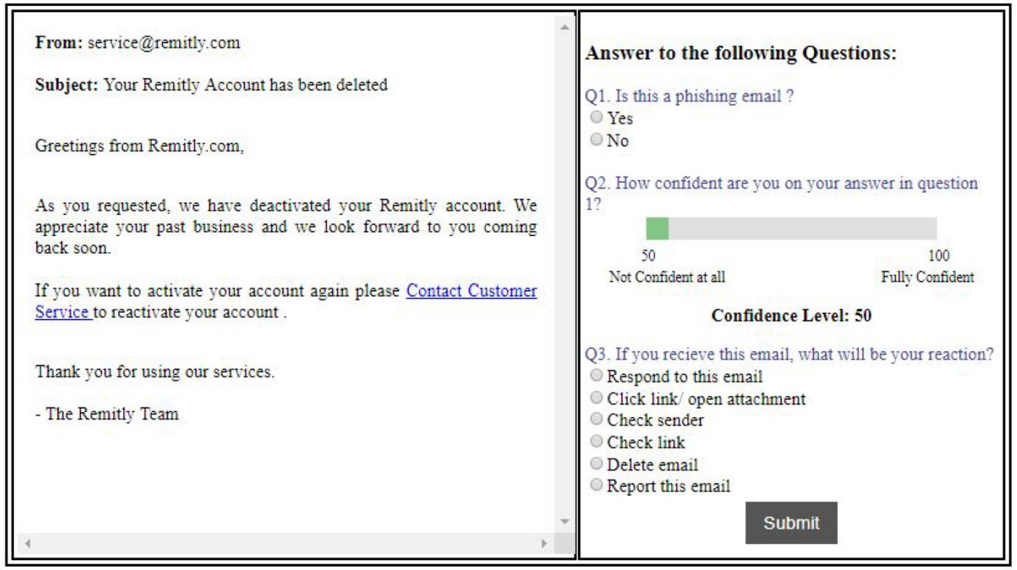
[2]Lebiere, C. (1999). A blending process for aggregate retrievals. In *Proceedings of the 6th ACT-R Workshop*. George Mason University, Fairfax, Va.

[3]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.

Building a generalizable IBL model

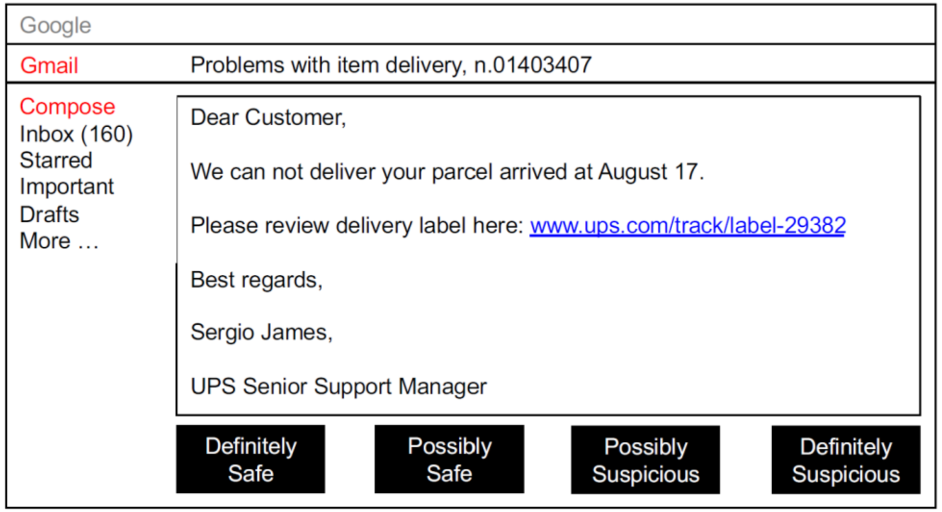
- Phishing Training Task (PTT)
  - 3 phases: Pre-test, Training, Post-test
  - 60 emails total, randomly selected according to frequency probabilities

Pre-test	Training	Post-test
10 trials	40 trials	10 trials
20% phishing	**25%, 50%, or 75%** phishing	20% phishing



PTT interface; from Fig. 1, Singh et al. (2019)

- Phishing Email Suspicion Test (PEST)
  - 4 types of emails:
    - Real-Phishing
    - Real-Ham
    - Simulated-Phishing
    - Simulated-Ham
  - Randomly presented 40 of each type in single testing phase
  - Generate rating of suspiciousness instead of classification



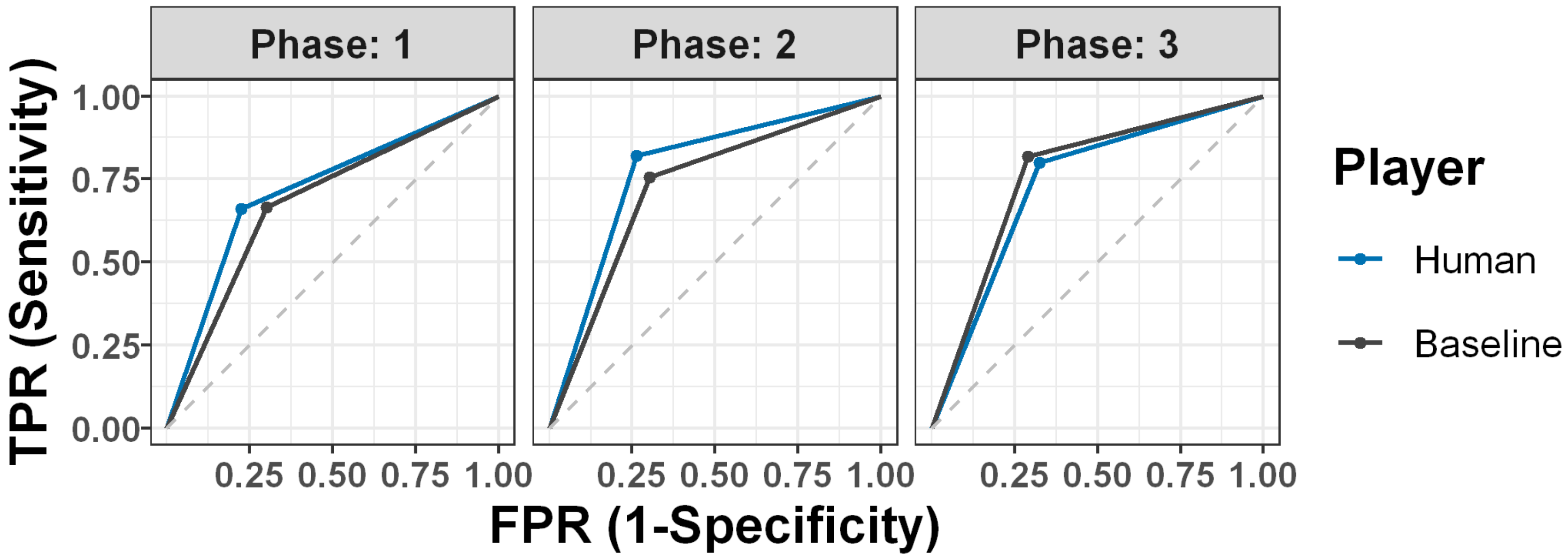
PEST interface; from Fig. 1, Hakim et al. (2020)

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.

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.*

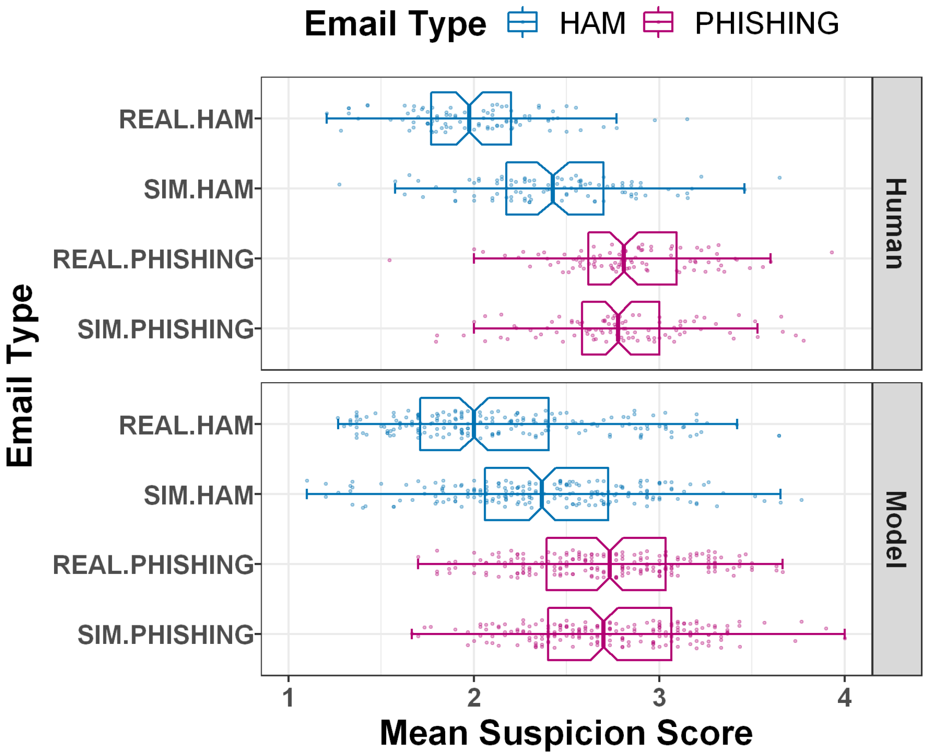
Model results of the PTT

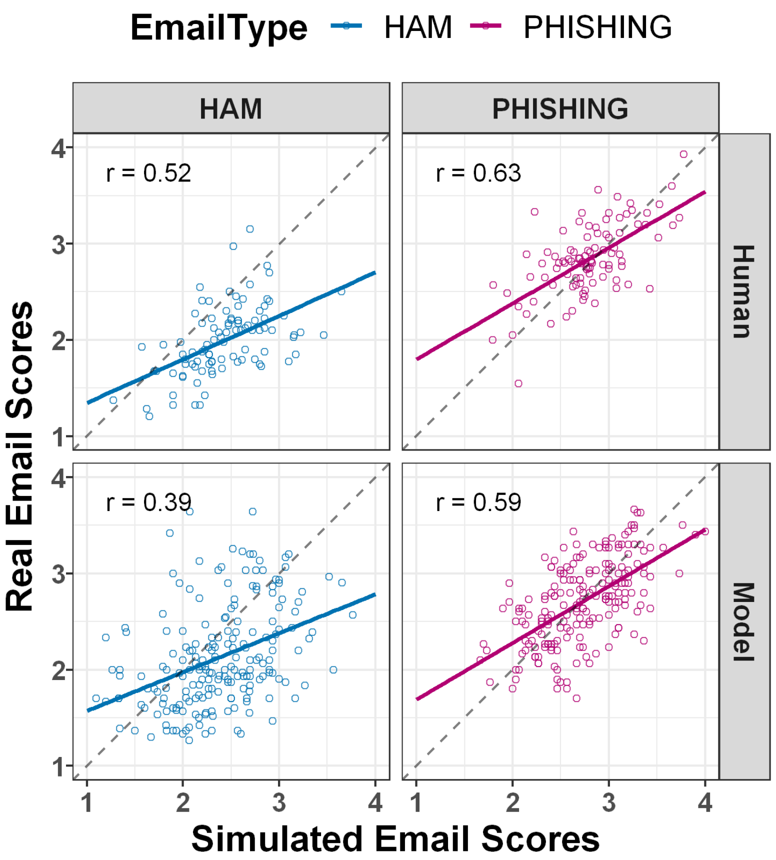
- Model accurately predicts end-user phishing discriminability and learning across the three phases of the experiment
  - Receiver Operating Characteristic (ROC) curves show that, like humans, model has difficulty distinguishing between ham and phishing emails, even after extensive training



Model results of the PEST

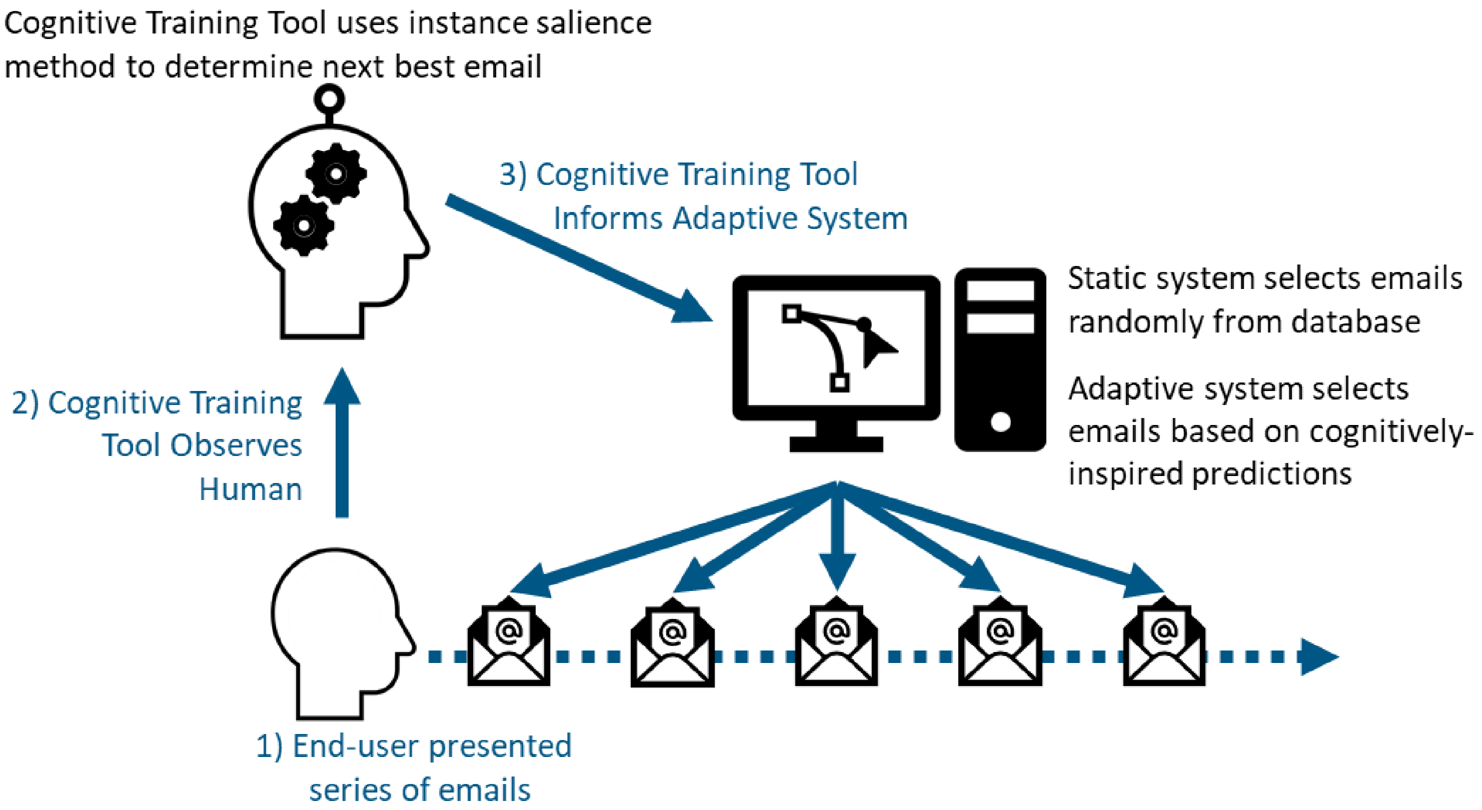
- Model accurately predicts individual differences of end-users in terms of rating real and simulated, ham and phishing emails on a scale of suspiciousness
  - Model shows greater variability due to running 300 simulated participants compared to only 97 humans





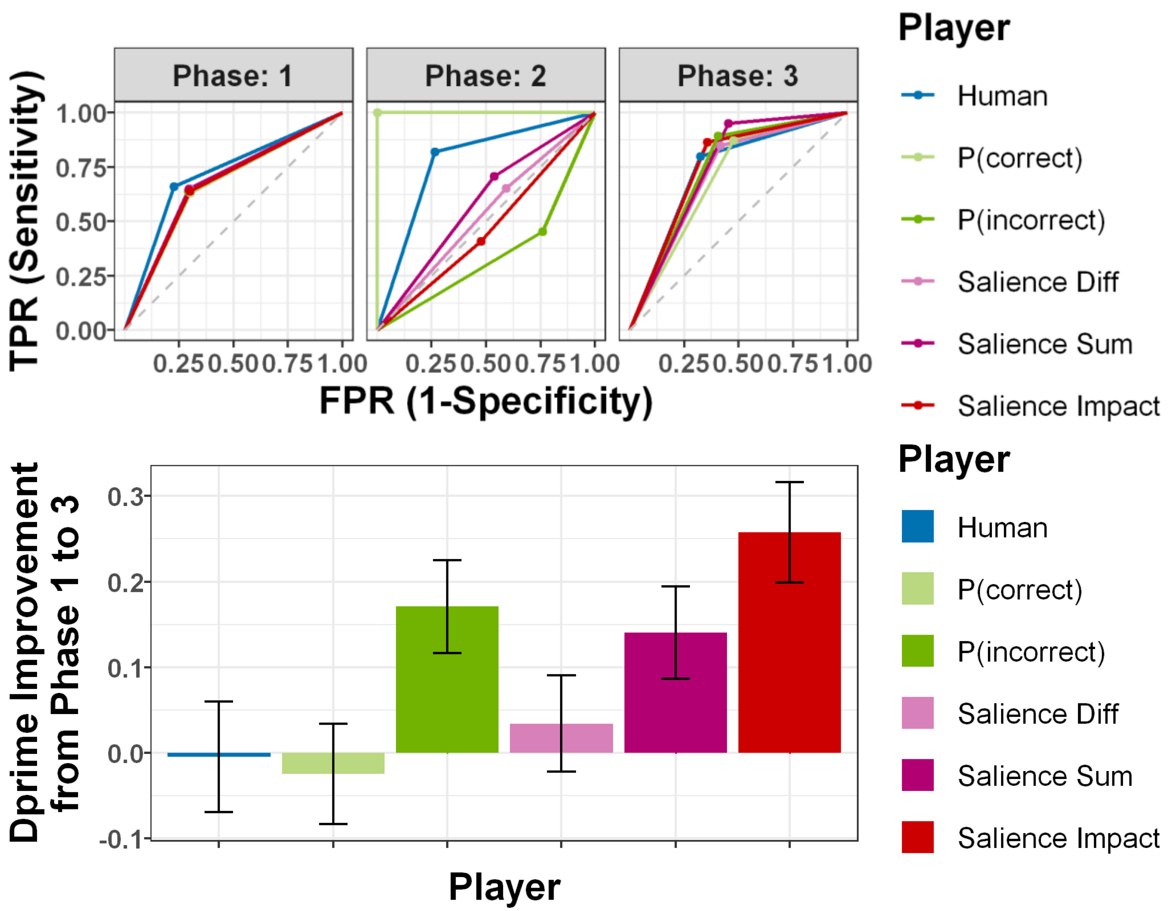
Personalized anti-phishing training

- Combines model-tracing techniques (e.g., used by cognitive tutors) and IBL cognitive modeling to predict human behavior and inform the Cognitive Training Tool
  - Requires little-to-no training data to make accurate predictions
  - Adapts to human decisions/experience
- Instance Salience computed to determine relative influence instances have on the decision
  - Derivative of blending equation
    - $S(V, A_k) = \frac{\partial V}{\partial A_k}\Big]_{V=V_o}$
  - Guides selection of best email to maximize discriminability
  - Goal is to make boundaries between categories more distinct in memory
    - Based on cognitive principles such as recency and frequency of instances, and their effects on the availability of information during retrieval processes



Model predictions of personalized training

- Human performance under static training methods compared to model predictions under 5 iterations of personalized training method
  - 2 methods based on estimated retrieval probabilities
    - ***P(correct)*** – selects email most likely to be classified correctly, based on estimated retrieval probabilities
    - ***P(incorrect)*** – selects email most likely to be classified incorrectly, based on estimated retrieval probabilities
  - 3 methods based on instance salience
    - ***Salience Diff*** – selects email with greatest absolute difference between the most salient in-category instance and out-category instance
    - ***Salience Sum*** – selects email with greatest absolute sum of saliences across all instances
    - ***Salience Impact*** – selects email that is most salient in their own category and least salient in the other category
      - Selected instance maximizes difference between the absolute value of the sum over the other probes of its own category and the absolute value of the same sum for the other category



Limitations

- In current experiment, database of phishing emails are highly similar to ham emails in terms of semantics
  - Also lacks context and knowledge of end-user interests and past experience with emails
  - Results could be better in a real-world situation if model is given a short history of an end-user’s experience with past emails and their interests
  - Model could perform better given additional cues beyond solely relying on semantics
    - Research shows that teaching end-users to identify relevant features can further improve discriminability
      - e.g.,
        - link/sender mismatches
        - appeals of urgency
        - offers of rewards
        - requests of credentials
      - Singh et al., 2020

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.

Conclusions

- Results highlight generality of model by predicting behavior across different tasks with different dataset
- Phishing susceptibility can be modeled as decisions from experience
  - Semantic similarity between email features useful for generating accurate predictions
    - Provides an automated process for generating similarities that allows for adaptable cognitive models
    - Future anti-phishing training should be geared toward training end-users to detect high-level, expert features
- Our automated cognitive training system is expected to contribute to savings in training personnel and time needed for training, and to improve overall security from threats of phishing emails by empowering end-users with the ability to be pro-active in defense against phishing attacks
- Human experiments under way to validate effectiveness of personalized training
- Broad applications
  - In other research, applying instance salience technique to Intrusion-Detection

Questions?

cranford@cmu.edu

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.