12-12-2021

Developing a measure of adversarial thinking in social engineering scenarios

Justin Scott Giboney
Brigham Young University, justin_giboney@byu.edu

Ryan M. Schuetzler
Brigham Young University

G Mark Grimes
University of Houston

Developing a measure of adversarial thinking in social engineering scenarios

Justin Scott Giboney[1]
Information Systems, Brigham Young University,
Provo, Utah, USA

Ryan M. Schuetzler
Information Systems, Brigham Young University,
Provo, Utah, USA

G. Mark Grimes
Decision and Information Sciences, University of Houston,
Houston, Texas, USA

ABSTRACT

Social engineering is a major issue for organizations. In this paper, we propose that increasing adversarial thinking can improve individual resistance to social engineering attacks. We formalize our understanding of adversarial thinking using Utility Theory. Next a measure of adversarial thinking in a text-based context. Lastly the paper reports on two studies that demonstrate the effectiveness of the newly developed measure. We show that the measure of adversarial thinking has variability, can be manipulated with training, and that it is not influenced significantly by priming. The paper also shows that social engineering training has an influence on adversarial thinking and that practicing against an adversarial conversational agent has a positive influence on adversarial thinking.

Keywords: Adversarial thinking, social engineering, measure development, conversational agents, chatbots.

INTRODUCTION

[1] Justin Giboney. justin_giboney@byu.edu +1 801 422 4285

Social Engineering (SE), in its broadest definition, is the act of "skillfully maneuvering human beings to take action in some aspect of their lives" (Hadnagy 2010, p. 10). SE is an attack vector that takes advantage of human habits and biases to gain unauthorized access to resources or information. SE techniques are employed to circumvent technical security measures, exploit human decision-making, and obtain access to confidential information or systems from businesses, organizations, or individuals (Abbasi et al. 2021). The Verizon Data Breach Investigation Report identified 3,841 SE incidents in 2020 (Verizon 2021). SE breaches cost companies and countries billions of dollars every year (Mouton et al. 2017). As a result, companies dedicate resources developing policies and training employees to comply with those policies. For example, many organizations conduct internal phishing attacks to evaluate employee's understanding of good security practices (Wright and Thatcher 2021).

Adversarial thinking (c.f., Hamman and Hopkinson 2016) is the ability to reason about the actions and goals of a malicious actor. Adversarial thinking proponents believe that if a person can reason like an attacker, they will be better at defending themselves from the attacker (Thompson et al. 2018). While there has not been much theoretical development of adversarial thinking, training on Game Theory has been used to improve people's ability to reason about attacker behavior. This paper will further develop the theoretical tenets of adversarial thinking and develop a measure of adversarial thinking in a social engineering context.

The remainder of this paper describes the development and testing of a measure of adversarial thinking in SE. The measure was created following accepted protocols (Mackenzie et al. 2011) used in cybersecurity measure development (Giboney et al. 2016) and is based on the theoretical concept of Utility Theory (Fishburn 1968). We provide an overview of the theory and literature used to develop the measure. Then we report on the findings from two data collections

to validate the accuracy of the measurement. Finally, we discuss the implications of this measure for researchers and practitioners.

SCALE DEVELOPMENT

MacKenzie et al. (2011) recommend six phases for scale development: 1) conceptualization, 2) development of measures, 3) model specification, 4) scale evaluation and refinement, 5) validation, and 6) norm development. The following sections will discuss each of the phases in the measurement.

Conceptualization

During the conceptualization phase the researchers provide a clear definition of the construct of interest and provide theoretical tenets of the construct that distinguish it from previously published constructs (Mackenzie et al. 2011). While this paper does not invent the concept of adversarial thinking, it attempts to further the theoretical development of the construct using Utility Theory (Fishburn 1968).

Thompson et al. (2018) define adversarial thinking as the "reasoning about actions and goals in a context in which there might be malicious actors attempting to defeat those goals and carry out their own nefarious actions" (p. 1). Hamman and Hopkinson (2016) define adversarial thinking as, "the ability to embody the technological capabilities, the unconventional perspectives, and the strategic reasoning of hackers" (p. 11). These definitions and other research suggest that cybersecurity adversarial thinking involves the understanding of the objectives and capabilities of malicious actors (c.f., Zoto et al. 2018), who bad actors are, their resources, access, their risk tolerance as well as knowing your own computer systems and their vulnerabilities (Sherman et al. 2017). However, it is more than just thinking about the attacker, it is also thinking about an attacker that is thinking about a defender. Adversarial thinking is

loosely supported by game theory and utility maximization (Hamman et al. 2017; Schneider 2013). To our knowledge, the benefits of adversarial thinking have not been objectively measured, and more work is needed in this area (Dark and Mirkovic 2015).

Utility Theory assumes that there is value in products and decisions. Utility is the amount of benefit (satisfaction or value) someone receives from a product, service, or taking an action (Fishburn 1968). Utility can be mathematically modeled as a curve. The top of the curve represents the marginal benefit of taking an action.

In a cybersecurity context, companies must think about the utility of their actions, and particularly the marginal utility of additional money spent to secure their systems. For example, spending $1,000 to improve the security of a network could reduce the risk of network compromise by 20%. Spending an additional $1,000 might decrease the risk by only another 10%. We can measure the total utility someone receives by calculating the area under the curve from their starting location (before taking an action) to their ending location (after taking an action) (Stigler 1950). Integrating from 0 to $2,000 using a function gives us a total 30% reduction. However, the value of money spent on cybersecurity is not constant from $0 to $2,000 but is typically represented as a curve where the variable is the number of dollars spent on a particular part of cybersecurity.

Also, within the context of securing a system, there is rarely only one thing on which effort can be spent to improve cybersecurity. Security teams need to consider all their cybersecurity functions to determine how best to allocate scarce resources to provide maximum utility. If $1,000 spent on phishing prevention has higher utility than the same amount spent on network security, Utility Theory indicates that a rational organization should choose the action

(or set of actions) that maximize the value of the money spent (Fishburn 1968), in this case phishing prevention.

Utility Theory considerations model human behavior where players act to maximize their expected utility where they have subjective beliefs about the probability of another's actions (Banks et al. 2011). To calculate utility for either player in these scenarios, we need to use the pairs of actions, one from each party. In a scenario with imperfect knowledge, such as in real-world cybersecurity decisions, adversary actions are uncertain variables that must be used to try to predict the utility of various actions (Naveiro et al. 2019; Rios Insua et al. 2009).

When modeling cybersecurity impacts and decision making there are many factors to consider including: cost to the organization, cost to other organizations, harm to people, and even environmental damage (Couce-Vieira et al. 2020). The cost and probability are often termed likelihood and impact in frameworks like NIST's Risk Management Framework when determining risk (Derbyshire et al. 2021). In many risk calculations, the likelihood of success of an attack consider the probability of success and costs such as time and needed financing leading to a probability density function used in mathematical modeling (Derbyshire et al. 2021).

Adversarial thinking is founded in the idea that people do not understand or are not aware of cybersecurity risks or actions of attackers. Many studies have shown that people do not understand cybersecurity risks (McShane et al. 2021). Utility (e.g., successfully preventing an attack) is derived from the costs and probability of an action. We propose a more formal modeling of adversarial thinking with the following propositions:

> P1. An understanding of the costs of an attacker's actions is a component of adversarial thinking.
> P2. An understanding of the probability of success of an attacker's actions is a component of adversarial thinking.

> P3. An understanding of the utility of an attacker's actions is a component of adversarial thinking.

As a cybersecurity professional masters the concepts of adversarial thinking, they will be better able to predict the costs, probability, utility of the actions that maximize utility of the attacker. Doing so will make them better equipped to optimize the allocation of resources to defend against attacks. It is not as simple as "What would an attacker do?" Instead, adversarial thinking involves predicting attackers' actions based on what an attacker might be expected to know about the company's defenses. If an attacker knows about certain defenses, their tactics will change, affecting the calculus of how defense funds could be spent most effectively. This leads to our last proposition:

> P4. As a person increases in adversarial thinking, they are better able to defend themselves and their organization.

Development of measures

Currently there are only a few limited measures of adversarial thinking. Most are limited to the scenario in which security is being taught. A few studies (Hamman et al. 2017; Katz 2019) measured it with a log analysis resource allocation scenario. The scenario indicates that a company collects customer data on a knowingly insecure server. At the end of the work week the company moves the data to a secure server. During the week, the company allocates human log analysts to inspect the customer data on the insecure server looking for breaches. It is assumed that the more hours spent on a particular day investigating logs will increase the likelihood of catching a breach. For an attacker, it is assumed that the reward from breaching the insecure server increases linearly throughout the week. Katz asked students to assume the role of an attacker and asked which day of the week they would attack. Nearly half of the attackers chose to attack on Wednesday, with 36% on Tuesday, and 18% on Thursday. This provided the authors

with an attacker mentality. Later the authors approached another set of participants to play the defensive side by asking them to allocate a 40-hour work week across the five workdays. Comparing the allocation to the attacker profile gives insight to how much like an attacker the participant is thinking.

We took a similar approach to measure adversarial thinking in an SE context. We constructed a scenario in which "attacker" participants were instructed to social engineer a chatbot to obtain a password. We took four conversations from a pilot study of this bot as examples of attacker behavior. In each of these conversations, the participant successfully retrieved a password from the CA. We cut these conversations short, mixed in some statements from other conversations that fit with the flow of the original conversation and asked participants which statement came next. As participants order the statements, we can see how well participants match their thinking with those of attackers. We have included the conversations in **Error! Reference source not found.**. While we randomized the order of the statement choices for evaluation, here the first statement in each column represents what the attackers said. As participants order the possible statements, the closer the participant puts the correct statement to 'most likely' the better they score on the adversarial thinking measure.

Table 1. Conversations for measure of adversarial thinking in social engineering.

Conversation 1	**Conversation 2**	**Conversation 3**	**Conversation 4**
Thanks for chatting with RipTech, my name is Janet, how can I help you?	Thanks for chatting with RipTech, my name is Janet, how can I help you?	Thanks for chatting with RipTech, my name is Janet, how can I help you?	Thanks for chatting with RipTech, my name is Janet, how can I help you?
I would like to change my password	I've ran into some trouble with my password	Hello, Janet! My name is George, but my friends call me Georgie. I am super busy today and I don't have much	Hi Janet, i cant remember which email i used to sign up for, can you help me with this?
I'm not allowed to send out passwords like this, have the	I'm not allowed to send out passwords		

account owner contact me. Lina asked me to change it for her Like I said, I can't just give you the password info.	like this, have the account owner contact me. Who can send it to me? Listen, I would love to help but I really can't send it to you.	time, but I need your help. I understand you need my help quickly, but I can't break policy to help. I'm sorry, I'm just trying to get this done quickly. I don't want you to get in trouble with your boss. I follow the same rules as everyone else in the department, personal information can't be sent over chat.	I'm not allowed to send out passwords like this, have the account owner contact me.
Lina asked me to do it for her since I'm her assistant. I will be fired if I can't change it for her.	Nobody will know	That's fine, but I'm not a normal customer, I'm George Hughman, the CTO.	My wife is the account owner and she is busy taking car of our baby right now but we need to get access today because we are moving out of state in the next few days
I completely understand, is there any way you could help me some other way	How can you send out information?	What do you mean?	Can you please send which email we used to sign up?
Who can send it to me?	How can I get it?	How would I go about fixing my password?	I know but changing my password hasnt been working! i tried resetting it through my email but it hasnt worked yet
Lina asked me to get the account information from you	I really appreciate knowing my account is safe. What can I do	What can you do for me?	Where can I go to get the password?

to receive my password?

To allow for people to practice social engineering techniques, we designed a CA to act as a technical support representative of a fictitious company. The CA was first pilot tested in a capture-the-flag event, in which one flag was the Chief Technology Officer's (CTO's) password. Participants in the event were tasked with using SE techniques to extract the password from the tech support conversational agent. Some information about the CTO was available on the fictitious company's website so the attackers could use that information in the SE attacks to increase the credibility of their requests.

The CA uses Rasa, a machine learning chatbot development platform. When users send a message, Rasa classifies the message to a predefined intent. Our CA includes basic functionality to greet the user and respond to small talk. It also has eight intents modeled after common social engineering tactics: urgency, threatening, helping, reciprocity, consensus, context, authority, and fear of missing out (FOMO). The intents are then placed into stories that are used to manage the flow of conversation. Our stories center around users employing different social engineering tactics to persuade the CA to give up the password.

The chatbot uses an internal trust metric to determine when users have demonstrated enough social engineering techniques to merit receiving the password. As users successfully employ the common social engineering techniques, trust increases to a threshold, after which the chatbot responds by "reluctantly" giving the password and requesting follow-up from the CTO. For example, when a user states that he is Lina's secretary and that Lina has a meeting in 10 minutes and needs her password, the user is employing an urgency technique. The CA has been trained to recognize urgency messages and the user's trust score is increased by 25. After a few

more messages that use social engineering techniques, the user's trust score increases above the 70 threshold and the chatbot responds with the password.

We asked participants to read the conversation and rank order the four statements in terms of most likely to least likely. The statements were presented to users in a random order. In this way we could see how likely they think the statement from the attacker is part of the actual conversation.

Scale evaluation – Study 1

Participants we recruited from a Junior-level cybersecurity course at a university in the United States. Students had already been introduced to SE earlier in the semester. Eight females and 37 males participated in the study. This proportion represents the proportion of students enrolled in the course.

In study 1 we wanted to validate our measure of adversarial thinking and view the effect of priming on the measure. We wanted to see if the mention or training of social engineering techniques influenced the adversarial thinking measurement. Students were randomly chosen to either receive social engineering training in the form of an infographic or no training and randomly chosen to be told that the conversations they were about to receive were from a social engineer (i.e., 2x2 experimental design). There we 11 in the 'no training, no priming' condition, 10 in the 'priming only' condition, 13 in the 'training only' condition, and 11 in the 'training and priming' condition.

The first thing we wanted to see is whether the conversations moved together to create multiple measures of the same construct (i.e., adversarial thinking). To test this, we ran an ANOVA to see if any of the conversations reported a different score than the others. The ANOVA was significant (F-value = 5.685; p = .001), so we ran a set of pairwise comparisons.

The pairwise comparisons showed that conversation 2 was different than 1 ($p = 0.0014$) and 4 ($p = .0073$). After dropping conversation 2, we ran a Cronbach's alpha test which had a higher standard alpha (.62) than any if an item were dropped. This left us with three conversations that we could use as a measure of adversarial thinking. We averaged the position across the three conversations to create a composite score of adversarial thinking for each participant. Using the new adversarial thinking measure, we ran a linear model to check the effects of training and priming on the measure. In the linear model the effect of training alone was significant ($p < .05$) while the effect of priming and the interaction between training and priming were not significant.

Norm development – Study 2

The final step in scale development is to understand how the scale works on a broader scale. To help develop a norm for the scale we ran a survey using the develop SE chatbot and an established conversational agent named ELIZA. ELIZA plays the role of a Rogerian psychotherapist (Weizenbaum 1966). ELIZA was one of the first well-known conversational agents and is considered a source of comparison for other conversational agents.

Participants we recruited from Amazon's Mechanical Turk (MTurk) participants in the United States. Using MTurk allowed us to reach a broader demographic than available at the university. We recruited 99 participants from MTurk with a pay of $4 (about $16 per hour). 15 of them did not pass an attention check in the survey. Of the remaining 84 participants, 39 reported they were male, 44 reported they were female, and 1 preferred not to identify. The ages ranged from 27 years to 69 years with the mean age being 42.8 years old with a standard deviation of 9.9 years.

For study 2 we conducted another 2x2 manipulation of training x chatbot. Participants were presented with a training prior to interacting with a chatbot. That training was either a

training about SE and its related techniques, or an unrelated training about conversational agents and their history. For the chatbot manipulation, participants interacted with either the SE victim chatbot or with ELIZA. Through this we aimed to isolate the effects of training and chatbot interaction on adversarial thinking.

There were 20 participants in the SE victim with CA training condition, 26 in the SE victim with SE training condition, 21 in the ELIZA with CA training condition, and 17 in the ELIZA with SE training condition. 14 of the participants had a major issue (e.g., the CA stopped responding) during their interaction with the CA. We removed these from the analysis.

We used all four conversations from our measure of adversarial thinking. We did this to have a second check that the second conversation would not work. We again ran an ANOVA to check for differences in the conversations. The ANOVA was significant (F value = 38.97, $p < 0.001$). We again ran a pairwise comparison and showed that, again, conversation 2 was statistically different than the other three conversations. However, this time, conversation 4 was also statistically different than the other three conversations. This left us with two conversations to create the measure of adversarial thinking. We averaged the position of the remaining two conversations.

We ran a linear model to test the training type (SE or CA), the CA (ELIZA vs SE CA), and the interaction between the two. The type of training ($p = 0.003$) and the CA ($p = 0.019$) were both significant, but the interaction was not ($p = 0.933$). When participants received SE training their adversarial thinking was 1.59 compared to 2.67 when they received CA training (where a lower score indicates a higher level of adversarial thinking). When the participants interacted with the SE CA their adversarial thinking was 1.70 compared to 2.51 when interacting with ELIZA. See Figure 1.

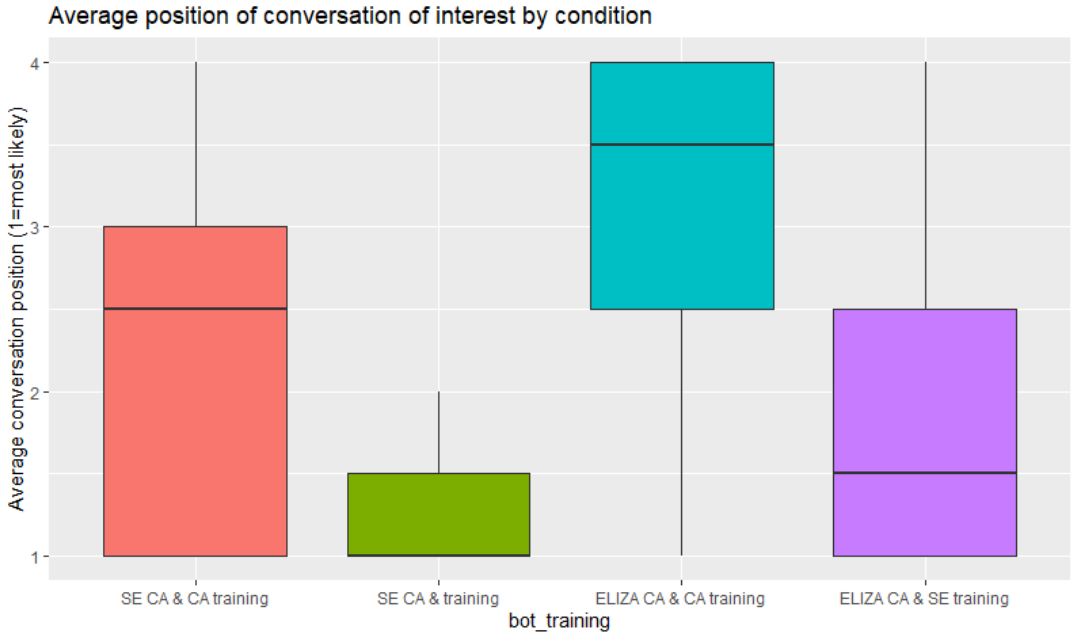

Figure 1. Results of adversarial thinking across conditions.

CONCLUSION

In summation, this paper reports on the formalization of adversarial thinking by introducing four propositions based on Utility Theory. Secondly, this paper introduces a mechanism to measure adversarial thinking in a social engineering context using a conversational agent. Lastly, this paper reports the results of two studies. We learned from study 1 that we have a measure of adversarial thinking in a SE chat scenario that has variability and can be manipulated with training that is not influenced significantly by priming. In study 2 we learned that social engineering training has and influence on adversarial thinking. Secondly, we learned that practicing against an adversarial CA also had a positive influence on adversarial thinking.

REFERENCES

Abbasi, A., Dobolyi, D., Vance, A., and Zahedi, F. M. 2021. "The Phishing Funnel Model: A Design Artifact to Predict User Susceptibility to Phishing Websites," *Information Systems*

Research (32:2), pp. 410–436. (https://doi.org/10.1287/isre.2020.0973).

Banks, D., Petralia, F., and Wang, S. 2011. "Adversarial Risk Analysis: Borel Games," *Applied Stochastic Models in Business and Industry* (27:2), pp. 72–86. (https://doi.org/10.1002/asmb.890).

Couce-Vieira, A., Insua, D. R., and Kosgodagan, A. 2020. "Assessing and Forecasting Cybersecurity Impacts," *Decision Analysis* (17:4), pp. 356–374. (https://doi.org/10.1287/DECA.2020.0418).

Dark, M., and Mirkovic, J. 2015. "Evaluation Theory and Practice Applied to Cybersecurity Education," *IEEE Security and Privacy* (13:2), IEEE, pp. 75–80. (https://doi.org/10.1109/MSP.2015.27).

Derbyshire, R., Green, B., and Hutchison, D. 2021. "'Talking a Different Language': Anticipating Adversary Attack Cost for Cyber Risk Assessment," *Computers and Security* (103), Elsevier Ltd, p. 102163. (https://doi.org/10.1016/j.cose.2020.102163).

Fishburn, P. C. 1968. "Utility Theory," *Management Science* (14:5), pp. 335–378.

Giboney, J. S., Proudfoot, J. G., Goel, S., and Valacich, J. S. 2016. "The Security Expertise Assessment Measure (SEAM): Developing a Scale for Hacker Expertise," *Computers & Security* (60), Elsevier Ltd, pp. 37–51. (https://doi.org/10.1016/j.cose.2016.04.001).

Hadnagy, C. 2010. *Social Engineering: The Art of Human Hacking*, John Wiley & Sons.

Hamman, S. T., and Hopkinson, K. M. 2016. "Teaching Adversarial Thinking for Cybersecurity," *Journal of The Colloquium for Information System Security Education (CISSE)* (June), pp. 93–110.

Hamman, S. T., Hopkinson, K. M., Markham, R. L., Chaplik, A. M., and Metzler, G. E. 2017. "Teaching Game Theory to Improve Adversarial Thinking in Cybersecurity Students," *IEEE Transactions on Education* (60:3), IEEE, pp. 205–211. (https://doi.org/10.1109/TE.2016.2636125).

Katz, F. 2019. "Adversarial Thinking: Teaching Students to Think Like a Hacker," *KSU Proceedings on Cybersecurity Education, Research and Practice* (10), p. 55. (https://digitalcommons.kennesaw.edu/ccerp%0Ahttps://digitalcommons.kennesaw.edu/ccerp/2019/education/1).

Mackenzie, S. B., Podsakoff, P. M., and Podsakoff, N. P. 2011. "Construct Measurement and Validation Procedures in MIS and Behavioral Research: Integrating New and Existing Techniques," *MIS Quarterly* (35:2), pp. 293–334.

McShane, M., Eling, M., and Nguyen, T. 2021. "Cyber Risk Management: History and Future Research Directions," *Risk Management and Insurance Review* (24:1), pp. 93–125. (https://doi.org/10.1111/rmir.12169).

Mouton, F., Teixeira, M., and Meyer, T. 2017. "Benchmarking a Mobile Implementation of the Social Engineering Prevention Training Tool," *2017 Information Security for South Africa - Proceedings of the 2017 ISSA Conference* (2018-Janua), pp. 106–116. (https://doi.org/10.1109/ISSA.2017.8251782).

Naveiro, R., Redondo, A., Ríos Insua, D., and Ruggeri, F. 2019. "Adversarial Classification: An Adversarial Risk Analysis Approach," *International Journal of Approximate Reasoning* (113), Elsevier Inc., pp. 133–148. (https://doi.org/10.1016/j.ijar.2019.07.003).

Rios Insua, D., Rios, J., and Banks, D. 2009. "Adversarial Risk Analysis," *Journal of the American Statistical Association* (104:486), pp. 841–854. (https://doi.org/10.1198/jasa.2009.0155).

Schneider, F. B. 2013. "Cybersecurity Education in Universities," *IEEE Security and Privacy* (11:4), IEEE, pp. 3–4. (https://doi.org/10.1109/MSP.2013.84).

Schuetzler, R. M., Giboney, J. S., Grimes, G. M., and Rosser, H. K. 2021. "Deciding Whether and How to Deploy Chatbots," *MIS Quarterly Executive* (20:1), pp. 1–15. (https://doi.org/10.17705/2msqe.00039).

Sherman, A. T., DeLatte, D., Neary, M., Oliva, L., Phatak, D., Scheponik, T., Herman, G. L., and Thompson, J. 2017. "Cybersecurity: Exploring Core Concepts through Six Scenarios," *Cryptologia* (42:4), Taylor & Francis, pp. 1–42. (https://doi.org/10.1080/01611194.2017.1362063).

Stigler, G. J. . 1950. "The Development of Utility Theory," *Journal of Political Economy* (58:4), pp. 307–327.

Thompson, J., Herman, G., Scheponik, T., Oliva, L., Sherman, A., Golaszewski, E., and Phatak, D. 2018. "Student Misconceptions about Cybersecurity Concepts: Analysis of Think-Aloud Interviews," *Journal of Cybersecurity Education, Research and Practice* (2018:1), p. 5. (https://doi.org/10.13016/M2XD0R18K).

Verizon. 2021. "2021 Data Breach Investigations Report." (https://enterprise.verizon.com/resources/reports/2021-data-breach-investigations-report.pdf).

Weizenbaum, J. 1966. "ELIZA–A Computer Program for the Study of Natural Language Communication between Man and Machine," *Communications of the ACM* (9:1), pp. 36–45.

Wright, R., and Thatcher, J. B. 2021. "Phishing Tests Are Necessary. But They Don't Need to Be Evil," *Harvard Business Review*. (https://hbr.org/2021/04/phishing-tests-are-necessary-but-they-dont-need-to-be-evil, accessed April 1, 2021).

Zoto, E., Kowalski, S., Frantz, C., Lopez-Rojas, E., and Katt, B. 2018. "A Pilot Study in Cyber Security Education Using CyberAIMs: A Simulation-Based Experiment," *IFIP Advances in Information and Communication Technology* (531:September), pp. 40–54. (https://doi.org/10.1007/978-3-319-99734-6_4).