Sponsor: Submission to the SCORE committee for review,
POCs Brad Martin (wbmarti@tycho.ncsc.mil), Celeste Paul (clpaul@tych.ncsc.mil)

Title of Proposal: Hacking Sunk Cost

PI: Robert Gutzwiller, PhD, Arizona State University (**ASU**)

Co-PI: Cleotilde (Coty) Gonzalez, PhD, Carnegie Mellon University (**CMU**)

Performers:
Palvi Aggarwal, PhD, Carnegie Mellon University (CMU)
Chelsea Johnson, Arizona State University (ASU)
Joe Gervais, Arizona State University (ASU)
Chris Kiekintveld, PhD, University of Texas El Paso (UTEP)

Lead Organization: Arizona State University (ASU)

Sub Organization: Carnegie Mellon University (CMU) – Period 1 and Option period
University of Texas, El Paso (UTEP) – Option period only

Technical Point of Contact: Dr. Robert Gutzwiller, Arizona State University, (480) 727-3716, robert.gutzwiller@asu.edu

Administrative Point of Contact: Heather Clark, (480) 727-4625, ASU.Awards@asu.edu

Estimated Cost:
Period 1 (12 mo) = $60,000
Option Period (12 mo) = $108,997

Duration of Project:
Period 1: 12 months from date of award
Option Period: 12 months following Period 1

Goal: To induce and exacerbate biases in cyber attackers to increase attacker cost.

Humans have a tendency to continue with specific strategy because of their prior investments, such as money, effort or time (Arkes & Blumer, 1985). Cyber attackers who invest a lot of resources in the attack process may tend to consider historical costs in assessing the value of a future outcome. Costs already spent should be evaluated as "sunk" and carry less weight than current or incremental investments (Thaler, 1980). However, when coping with uncertainty in decision-making, humans are vulnerable to biased thinking (Carter, Kaufman, & Michel, 2007). In the sunk cost fallacy, these vulnerabilities may lead to the decision to continue to work on a task (e.g. ongoing commitment) even though this decision is more costly. This fallacy relates to loss aversion in which the potential pain of losing a resource (for example, from quitting the task at hand) is greater than the potential pleasure of gaining a resource (Kahneman & Tversky, 1979).

Defenders could take advantage of sunk cost fallacy to make attackers do certain actions of their choice because the fallacy often operates outside of direct consciousness (Carter, Kaufman, & Michel, 2007). This means that even when people are presented with information that suggests the ongoing commitment is a poor choice, they may still choose to continue with the task and even increase or escalate their investment (Brockner et al., 1986; Schwenk, 1984).

Sunk cost is one of many different biases that are being explored as part of an ongoing project at Arizona State University (ASU) on Oppositional Human Factors (OHF). ASU has found evidence in red team behaviors for confirmation bias, anchoring, attentional tunneling and the use of various heuristics (Gutzwiller et al., 2018; 2019). The challenge is to intentionally create these biases, rather than simply observe them in the wild. Toward that end, we propose to (a) develop scenarios that could induce and exacerbate sunk cost bias in cyber attackers to increase their attack cost; (b) test hypotheses in human subject experiments; (c) integrate the findings of this experiment in broader cyber defense scenarios. This work will leverage ongoing work at ASU and Carnegie Mellon University (CMU) on studying cyber attackers.

This work is a collaboration across Universities to examine techniques beyond cyber deception, that use OHF for defender advantage. We will examine a specific cognitive bias and examine how to identify, induce, and exacerbate it in a cyberattack scenario. The Laboratory for Advanced Cybersecurity Research will act as a technical advisor (Kim Ferguson-Walter). Chris K. from UTEP will act initially as an advisor and will transition in the Option year to paid performer.

1. Technical Plan

A) Develop scenarios in HackIT to induce and exacerbate sunk cost:

Scenarios in which sunk cost fallacy may come into play in a cyber-domain must be examined. We will increase the effort and time required (perhaps resources) to get to a certain point in the scenario to determine whether the participant will continue along that same strategy due to the sunk cost or abandon for a different option.

We plan to develop above scenarios in HackIT and collect human data to identify the sunk cost fallacy. HackIT tool provides ability to construct networks of different sizes, ability to gather information about networks using Nmap scanning tool, exploit certain nodes in the network and

steal information. A few of the variations or variables we will explore to induce sunk cost include in priority order, but are not limited to:

- **Alter time spent for scanning results to be returned to them**
- **Create files of different value available for exfiltration, change download speeds**
- Provide attacker limited number of zero-days
- Change results of scans to provide misinformation
- Create long versus short chains of stealing passwords and accessing machines to achieve the goal
- Password protected file that required solving a cypher

We plan on initially exploring a variety of ways to induce sunk cost, but use 1 or 2 at most in an experiment.

A.1 – Goal: Experimental Scenario Development
This will be the first goal of the project. Scenario development will be based on the theory of sunk cost, input from cyber SMEs, and the limitations of realism and the HackIT platform.

A.2 – Goal: Develop Measurement Methods
The second task within the scenario development is to ensure that our measurement of the bias (e.g., whether it was induced or not) must be defined. Prior literature on the sunk cost fallacy has focused on microeconomic theory and rational decision-making (Arkes & Ayton, 1999; Beeler & Hunton, 1997; Hastie & Dawes, 2001; Kahneman et al., 1991; Schwenk, 1984, 1986; Shaanan, 1994; Sharp & Salter 1997; Staw, 1976, 1981; Williams, 1986). However, these are (1) largely group-level assessments, (2) focused on business or economics, and (3) have not been applied to cybersecurity situations. Our work will create sunk cost effects and measure them in the cyber scenarios of HackIT where individuals operate and are focused on cyber effects.

B) Conduct experiments with human subject experiments to test hypotheses:
Both the Dynamic Decision Making Lab (*Aggarwal*, CMU) and the Applied Attention Research Lab (*Gutzwiller*, ASU) have robust online and in person capabilities to conduct human subjects experimentation. Both universities have access to computer science and cybersecurity participants to test the hypotheses of this project. Furthermore, both are currently conducted human subjects research in cybersecurity and have Internal Review Board (IRB) approvals in place that can be quickly modified to accommodate these experiments.

ASU brings a world-class IRB to support our work. Internal metrics show that in 2019 between January and July, the ASU IRB has processed 836 protocols in an average of 13 days from initial submission to formal approval (and less than 5 days for approval of modifications and amendments across over 900 requests). PI Gutzwiller also has extensive experience with IRBs having served on IRB boards as a full member and has several successfully executed protocols to conduct human subjects research (HSR).

A significant output of this project is the findings of the experiment combined with a deep dive into how to induce a cognitive bias against an attacker. We will also attempt to integrate the findings into the broader understanding of OHF and assess its relatedness to realistic cyber

scenarios using our SMEs. We expect our results will be useful to a variety of different cyber situations.

2. Capabilities | Management Plan | Personnel

We bring together experts in the areas of cyber security, oppositional human factors, game theory, and cognitive modeling for human-centered cyber security research. We also leverage ongoing technology being used at CMU to model and study cyber-attacks. HackIT is a simulated cyber-attack environment developed by CMU to run HSR experiments without the need for expert participants. It was designed to investigate the human actor effects and strategies used around cyber deception for cyber defense.

ASU has been focused on oppositional human factors (OHF), which has been studying the use of human limitations and biases in cybersecurity to leverage human cognitive biases to increase attacker cost on a network.

We will supplement work on this project by regular communications between CMU investigators and ASU investigators. Primarily this will occur via telecommunications such as Skype or Zoom and led by the PI to facilitate interactions. However, given the related work, we will take advantage of any other in-person meetings and conferences to discuss this work and progress (e.g., Human Factors conferences).

2.1 Grant Performers:

PI: Robert Gutzwiller, ASU will focus on oppositional human factors (OHF), which is studying the use of human limitations and biases in cybersecurity to leverage human cognitive biases to increase attacker cost on a network. His background in psychological research methods and prior experience in this area will ensure a good experimental design.
Co-PI: Coty Gonzalez, CMU will focus on supervising and mentoring CMU performers in developing the HackIT simulation and participate in experimental design and review of the work.
Palvi Aggarwal, CMU will focus on developing scenarios in HackIT a simulated cyber-attack environment to run HSR experiments, leveraging her expertise and experience in modeling and studying cyber attacker cognition. Her background in psychological research methods and prior experience in this area will ensure a good experimental design.
Joe Gervais, ASU will leverage his cyber red team subject matter expertise (SME) in design and development.
Chelsea Johnson, ASU will aid in design and development of experiment based on her expertise in bias research and her ongoing dissertation work. Her background in psychological research methods and prior experience in this area will ensure a good experimental design.
Chris Kiekintveld, UTEP will advise in Period 1 on developing the experiment and lead discussions on Option year potential tasks such as developing cognitive models of bias.

2.2 Technical Advisor:

Kimberly Ferguson-Walter, Laboratory for Advanced Cybersecurity Research, will act as an unfunded technical advisor on this project.

3. Statement of Work | Cost | Schedule

3.1 Statement of Work (SOW)

3.1.1 Period 1 Statement of Work: [*12 months, beginning on date of award*]

- Develop an experimental design to induce sunk cost in the cyber environment of HackIT (ASU & CMU)
- Program HackIT for this paradigm (CMU)
- Gain IRB approval (ASU)
- Conduct a controlled experiment with HackIT to induce and measure bias (ASU)
- Write a report / journal article on the results, and present the work at an upcoming conference (targeting HICSS 2021)

3.1.2 Option Period

In developing the above ideas, we noted that two other major questions exist of interest; (1) what method should be used to signal deception, true or false, to potential attackers? The use of human subjects experiments and the HackIT platform may be a good first step to answer this question. Second, (2) we recognized that much of behavior is related to goals and preference; as attacking contains an exploratory element and the chance of discovery, we believe that more work needs to be done to understand basic "attraction" properties of network elements. The first step could be accomplished with HackIT before it would benefit from being scaled into more realistic participants with more cybersecurity expertise and into a more realistic environment.

(1) *What method is best to send a signal about deception?* Many papers using game theory for cyber deception focus on a signaling problem which provides a true or false signal. Human subjects experiments have shown that informing attackers that deception is in use can affect their behavior. What techniques are most effective in a cyber scenario for informing attackers of these extra defenses? An example of different techniques to test include providing obvious deception that is badly configured (in addition to well-designed deception), or providing a fake document for exfiltration whose contents reveal the deception. These different methods deserve empirical evaluation. We may be able to include these in an experiment in HackIT.

(2) The need to determine what is "attractive" to a specific attacker for more personalized deception? Different system attributes and different files accessed are easily tracked in HackIT. Goal recognition based on what is attractive is an important aspect that might be related to attacker profile, and useful for an efficient defense profile. While this is a difficult challenge in general, by constructing a variety of scenarios in HackIT and giving participants different goals, a few basic properties of behavior could be measured and established (e.g., when given a set of machines, and a given set of their properties, which machine is targeted first? Does changing 1 property change the choice?). These findings could be tested with more advanced participants or expanded into more realistic conditions in later work.

(3) The potential to model bias. We recognize that building on existing work done by CMU and UTEP there may be potential to develop cognitive models of the bias to use in future

predictions. UTEP will lead this part of the effort in ensuring our experiments would help develop or validate the model, and in developing the model itself.

Option Period SOW [*12 months, beginning on date of Period 1 completion*]

- Develop an additional experiment targeting signaling and/or attractiveness of elements in the cyber environment
- Program HackIT for this paradigm
- Conduct a controlled experiment
- Write a report / article on the results

3.2 Anticipated Schedule of Performance:

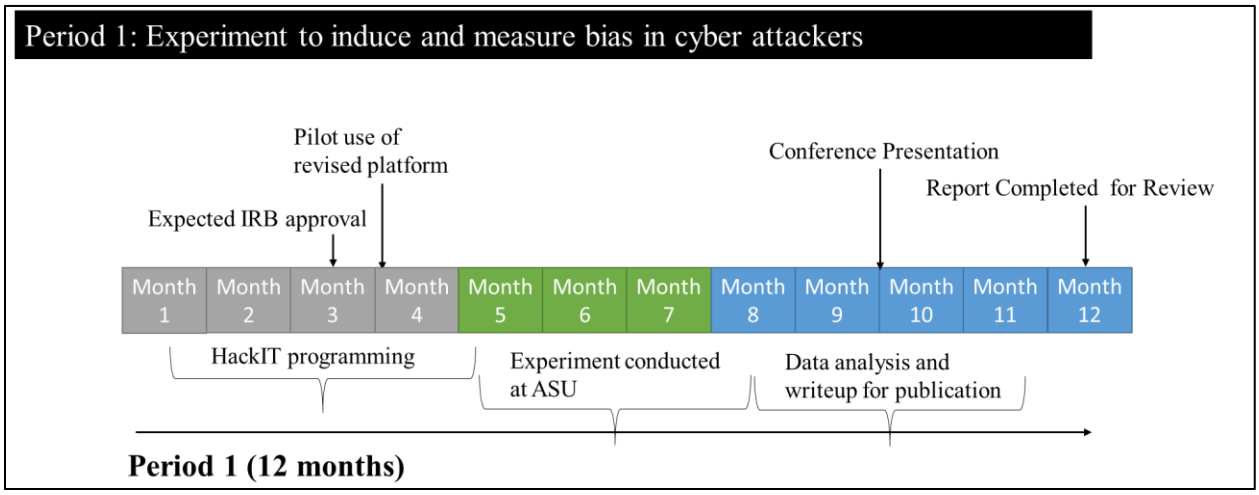

Figure 3.2.1 – Period 1 Schedule

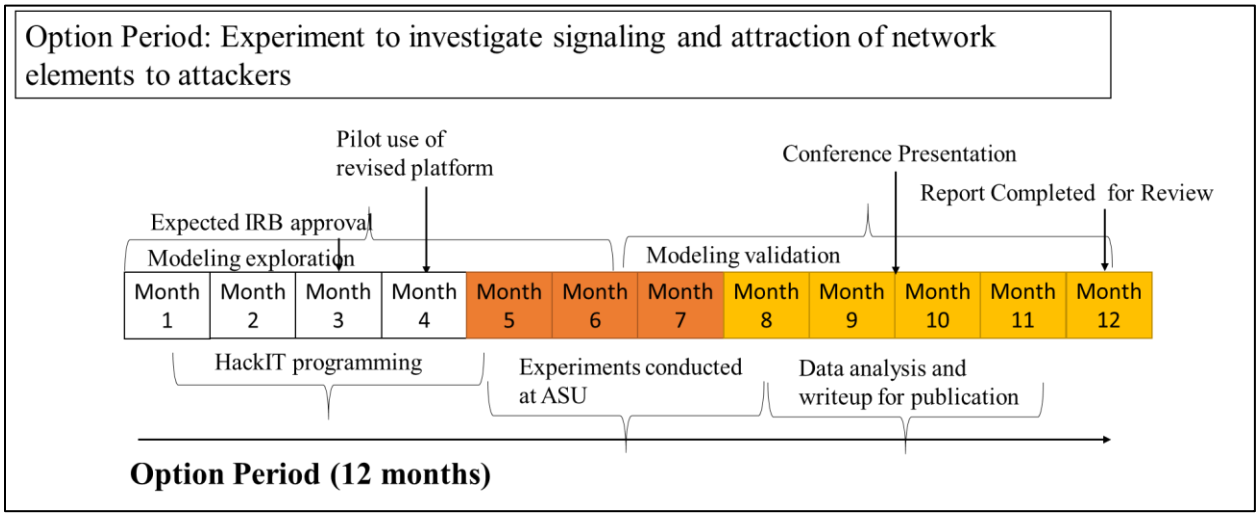

Figure 3.2.2 – Option Period Schedule

4.0 Related Work:

ASU is contracted to identify biases in red cyber attack data (as part of the OHF project). Work in this area will be leveraged to benefit the current proposal. Furthermore, Chelsea Johnson and Joe Gervais at ASU are graduate PHD students funded on OHF. Joe is a 30-year red team veteran. Together we will be working on biases in cyber for Johnson's dissertation, and will likely use this project to help develop additional work and explore the use of HackIT for OHF testing.

The Dynamic Decision Making Laboratory led by Prof. Gonzalez at CMU has developed a general research framework for the design of dynamic, adaptive and personalized deception strategies for cyber defense (Gonzalez, Aggarwal, Cranford, Lebiere, 2020). This research framework uses game theoretic algorithms of limited resource allocation and deceptive signaling (Cooney et. al, 2019a; Cooney et. al, 2019b) and cognitive models based on Instance Based Learning (IBL) (Cranford et. al, 2019; Cranford et. al, 2020) to develop adaptive and personalized cyber deception algorithms for stackelberg security games. These algorithms are tested using different interactive games that vary in complexity and realism; one such game is HackIT, developed by Aggarwal et al. (2019. HackIT is a generic web-based framework for cybersecurity to study human learning and decision-making of attackers and defenders. HackIT tool provides semantics such as network structure, network nodes with realistic features; deception tactics; and commands, which are used to communicate with the network. In this project, we plan to develop scenarios that induce or exacerbate the sunk cost fallacy in HackIT and learn the decision-making process of attackers.

The research group led by Dr. Kiekintveld at UTEP has several ongoing projects applying computational game theory to decision making for cybersecurity, including the optimizing strategies for using defensive deception techniques. This includes ongoing work in the area of behavioral game theory with collaborators in psychology and cognitive modeling, working to develop better predictive models of how human attackers make decisions, including their biases and mistakes. These models are integrated into game theoretic models which can calculate best response strategies for playing against the humans, and can be used to conduct laboratory experiments with human participants to validate the performance of the models. For example, in one project UTEP is modeling how attackers learn over time as a function of different defensive strategies for honeypot deployment.

5.0 Budget Estimation (Table 5.1)

Current All-Period Totals	Period 1 1/1/2020 12/31/2020	Period 2 1/1/2021 12/31/2021	Cumulative
Personnel:	**$3,381**	**$19,343**	**$22,724**
Travel:	**$2,000**	**$5,000**	**$7,000**
Human Subject:	**$1,200**	**$1,200**	**$2,400**
Subaward/Subcontract:	**$35,418**	**$60,287**	**$95,705**
Total F&A:	**$18,001**	**$23,167**	**$41,168**
Project Total:	**$60,000**	**$108,997**	**$168,997**

6.0 **References** (bolded names are investigators on this grant)

Aggarwal, P., Gautam, A., Agarwal, V., **Gonzalez**, C., & Dutt, V. (2019, July). HackIT: A human-in-the-loop simulation tool for realistic cyber deception experiments. *In International Conference on Applied Human Factors and Ergonomics* (pp. 109-121). Springer, Cham.

Aggarwal, P., **Gonzalez**, C., & Dutt, V. (2016). Cyber-security: role of deception in cyber-attack detection. In *Advances in Human Factors in Cybersecurity* (pp. 85-96). Springer, Cham.

Arkes, H. R., & Blumer, C. The psychology of sunk cost. (1985) *Organizational Behavior and Human Decision Processes*, 35(1), 124-140.

Arkes, H.R., & Ayton, P. (1999). The sunk cost and Concorde effects: are humans less rational than lower animals?, *Psychological Bulletin*, *125*, 591-600.

Beeler, J.D., & Hunton, J.E. (1997). The influence of compensation method and disclosure level on information search strategy and escalation of commitment, *Journal of Behavioral Decision Making*, *10*, 77-91.

Brockner, J., Houser, R., Birnbaum, G., Lloyd, K., Deitcher, J., Nathanson, S., & Rubin, J.Z. (1986). Escalation of commitment to an ineffective course of action: the effect of feedback having negative implications for self-identity, *Administrative Science Quarterly, 31*(1), 109-27

Carter, C. R., Kaufmann, L., & Michel, A. (2007). Behavioral supply management: A taxonomy of judgment and decision-making biases. *International Journal of Physical Distribution & Logistics Management*, *37*(8), 631–669.

Cooney, S., Vayanos P., Nguyen T. H., **Gonzalez**, C., Lebiere, C., Cranford E. A., & Tambe, M. (2019a). Warning Time: Optimizing strategies signaling for security against boundedly rational adversaries. *Proceedings of the 18th International Conference on Autonomous Agents and Multi Agents Systems.* AAMAS, 2019. May 13-17, 2019, Montreal, CA. (page 1892). ISBN: 978-1-4503-6309-9

Cooney, S., Wang, K. Bondi, E., Nguyen, T., Vayanos, P., Winetrobe, H., Cranford, E. A., **Gonzalez**, C., Lebiere, C., & Tambe, M. (2019b). Signaling just enough: Learning to find the Goldilocks zone to improve adversary compliance in security games. *28th International Joint Conference on Artificial Intelligence (IJCAI-19) - Workshop 10: Artificial Intelligence in Business Security* (AIBS). August 10-16, 2019, Macao, China.

Cranford, E. A., **Gonzalez**, C., **Aggarwal**, P., & Lebiere, C. (2019). Towards personalized deceptive signaling for cyber defense using cognitive models. *Proceedings of the Proceedings of the 17th ICCM.*

Cranford, E. A., **Aggarwal**, P., **Gonzalez**, C., Cooney, S., Tambe, M., & Lebiere, C. (2020). Adaptive cyber deception: cognitively informed signaling for cyber defense. *In Proceedings of the 53rd Hawaii International Conference on System Sciences (HICSS).* 10.

Gonzalez, C., **Aggarwal**, P., Cranford, E. A., & Lebiere, C. (2020). Design of dynamic and personalized deception: A research framework and new insights. *In Proceedings of the 53rd Hawaii International Conference on System Sciences (HICSS).* 10.

Gutzwiller, R. S**.,** Ferguson-walter, K., Fugate, S., & Rogers, A. (2018). "Oh, look, a butterfly!" A framework for distracting attackers to improve cyber defense. *Proceedings of the Human Factors and Ergonomics Society*, *62*, 272–276.

Gutzwiller, R. S., Ferguson-Walter, K. J., & Fugate, S. J. (2019). Are cyber attackers thinking fast and slow? Exploratory analysis reveals evidence of decision-making biases in red teamers. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *63*.

Hastie, R., & Dawes, R.M. (2001), Rational Choice in an Uncertain World. Sage, London.

Kahneman, D., Knetsch, J.L., & Thaler, R.H. (1991), The endowment effect, loss aversion, and status quo bias, *Journal of Economic Perspectives, 5*(1), 193-206.

Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk, *Econometrica, 47*(2), 263–292.

Schwenk, C. R. (1984). Cognitive simplification processes in strategic decision-making. *Strategic Management Journal*, *5*(2), 111-128.

Schwenk, C. R. (1986). Information, cognitive biases and commitment to a course of action, *Academy of Management Review*, *11*(2), 298-310.

Shaanan, J. (1994). Sunk costs and resource mobility: an empirical study, *Review of Industrial Organization*, *9*(6), 717-30.

Sharp, D.J., & Salter, S.B. (1997). Project escalation and sunk costs: a test of the international generalizability of agency and prospect theories, *Journal of International Business Studies*, *28*(1), 101-121.

Staw, B.M. (1976). Knee-deep in the big muddy: a study of escalating commitment to a chosen course of action, *Organisational Behaviour and Human Performance*, *16*, 27-44.

Staw, B.M. (1981). The escalation of commitment to a course of action, *Academy of Management Review*, *6*, 577-587.

Thaler, R. (1980). Toward a positive theory of consumer choice, *Journal of Economic Behavior and Organization*, *1*(1), 39–60.

Williams, R. (1986). Concorde and dissent: explaining high technology project failures in Britain and France, *Public Administration*, *64*(2), 242-244.