Adversarial Cognitive Engineering (ACE) and Defensive Cybersecurity: Leveraging Attacker Decision-Making Heuristics in a Cybersecurity Task

Chelsea K. Johnson, Richard W. Van Tassel, Andrew Rogers, Temmie Shade, Kimberly Ferguson-Walter
Laboratory for Advanced CyberSecurity Research
ckjohn5@radium.ncsc.mil

Abstract

The role of cyberspace continues to expand, touching nearly every aspect in our lives. Critical information, when stolen, can be devastating to a nation's people, economy, and security. To defend against this threat, it is essential to understand the human behind the attack. A first step in developing new defenses where human attackers are involved is obtaining valid and reliable human performance and decision-making data. These data can be procured through rigorous human science research that experimentally evaluates foundational theory and measures human performance. Taking the key concepts from behavioral economics, the game-based testbed, CYPHER, was specifically designed to test the occurrence of the Sunk Cost Fallacy across multiple decisions in an abstract cyber environment. Evaluating decisions made over a series of actions to catch a fictitious cyber thief, we analyze the effects of two antecedents (uncertainty and project completion) and resource expenditure. Our results show that irrespective of condition, significantly more participants unnecessarily wasted resources, demonstrating behavior consistent with the Sunk Cost Fallacy. These data provide a baseline upon which to build artificial intelligence algorithms for automated cyber defense. [1]

Keywords: Cognitive Engineering, cybersecurity, sunk cost fallacy, decision-making bias, heuristics

1. Introduction

Artificial Intelligence/Machine Learning techniques are a hot topic for many applications including data analytics, knowledge base management, communications, and entertainment. Computational models to advance automation may also be an avenue for improving cybersecurity for cyber defense in critical arenas like national security. To build effective and useful tools and systems, technology development must first understand the needs, strengths and limitations of human operators.

Drawing from the research on Oppositional Human Factors (Gutzwiller et al., 2018) to disrupt usability, we propose **Adversarial Cognitive Engineering (ACE)** to spotlight the focus on antagonistic methods to exploit mental *and* analytical processes. Strategically applied for cybersecurity, cyber defenders can deceive and misdirect attacker behavior by tactically leveraging the constraints of cognition. Research spanning decades over a vast array of domains recognize that eliciting decision-making biases effectively influences, for example, consumer purchasing, information consumption, group relationships, and career choice. However, effective application to cyber defense in ongoing and time-constrained cyber tasks is unknown. This is a critical area to investigate in support of more proactive cyber defenses because tactics to elicit decision-making biases aim to strike at the core resource supply necessary to carry out adversarial goals (Steingartner et al., 2021).

Decision-making heuristics are unconscious cognitive processes based on simplified mental models of the world. Built from pattern recognition and experience, these models allow for quick decision-making (Simon, 1955). This means decisions are generally efficient and effective (Arkes, 1991; Kahneman, 2011; Kahneman & Tversky, 1984; Klein et al., 1986; Slovic et al., 2002). However, using heuristics, people seek the adequate rather than the best solution (Ericsson & Charness, 1994; Kahneman & Klein, 2009). Environmental pressures of uncertainty and difficulty can lead to cognitive inefficiencies like errors in judgment or biases — the observable and measurable outcomes of heuristics (Tversky & Kahneman, 1981, 1992).

Cyber operators make decisions in critical settings that are complex, dynamic, and fraught with heavy consequences. Moreover, information in cyber environments is frequently incomplete and

[1] Content in this paper is based on the first author's dissertation. See (Johnson, 2022) for complete documentation.

URI: https://hdl.handle.net/10125/106495
978-0-9981331-7-1

ambiguous. Like all humans, cyber attackers rely on decision-making heuristics. For cybersecurity, these errors in judgment present the opportunity for cyber defensive strategies to leverage attacker cognition and affect adversarial goals and outcomes. From this defensive perspective, the complex decision space presented to cyber attackers elevates the importance of understanding the strengths and limitations of their cognitive processes (Aharoni et al., 2011).

In this paper we present the initial findings of the CYPHER experiment for proof of concept on whether cognitive limitations like decision-making heuristics and biases, can be influenced and measured in a cyber task. We suggest the baseline data on human behavior is a critical piece upon which to build automated systems. Ideally, these data will support the development of cognitive models, to include representative decision trees for humans and machine agents alike. In the following sections, we present the relevant background research on heuristics and bias, the CYPHER test bed and experiment, analyses and results. We then discuss overall take aways and offer lessons learned for future work.

2. Background

Recently researchers have begun to study human decision-making bias in the cyber operational domain (Gomez & Whyte, 2022; Knott et al., 2013; Mancuso et al., 2015; Mancuso et al., 2014) and more limitedly, as a defensive strategy against attackers (Ferguson-Walter et al., 2021). Beginning with the foundational research in behavioral economics, over the past several decades, the effects of hundreds of decision-making heuristics and biases have been investigated. Examples include, but are not limited to: business, marketing, health care, consumerism, and politics. Our work contributes to, and extends the foundational theory to cybersecurity for cyber defense with the goal to operationalize and experimentally elicit decision-making heuristics and measure the resultant bias in a multi-step adversarial cyber task. Following a comprehensive review conducted by cybersecurity and cognitive science subject matter experts (Johnson et al., 2020), the Sunk Cost Fallacy (SCF) was chosen because of the measurable effects on attacker resources and the potential benefit to strengthen cyber defense. To this end, we specifically developed a novel online platform, CYPHER, to test the appearance of the Sunk Cost Fallacy.

2.1. Sunk Cost Fallacy

The SCF is defined as the tendency to continue with a specific strategy due to prior investments of money, effort, or time (Arkes & Blumer, 1985) irrespective of the cost-benefit trade-off. Because the heuristics that lead to this bias operate on an unconscious level, cyber defense strategies, like ACE, may leverage the SCF to induce inaccurate thinking in attackers. An adversary, once inside a network, may make errors and squander resources because of ACE. That is, decision-making heuristics may lead to inefficient activities to reach a designated target. For example, this may be especially disruptive for attackers who have identified numerous high-value targets, yet waste resources by persisting in an inefficient route instead of choosing an alternate tactic to achieve their objective. Consequently, we hypothesize that even when presented with evidence that current activities are more expensive than an alternative, attackers place a higher value on resources already spent, or sunk. In doing so, the decision is made to continue investing in the current activity rather than switching to an alternate route that will satisfy the original purpose or goal. The SCF is demonstrated by the continued preference for the option whose costs have already been incurred.

There are varied explanations for why humans fall prey to the SCF. In this project, we focus on two important factors where the research is inconclusive: uncertainty and project completion (Friedman et al., 2007).

Uncertainty is defined as being in a situation where relevant knowledge is unknown and/or unavailable (Augier & Teece, 2021; Knight, 1921). In other areas of research, uncertainty has been shown to: (1) slow decision-making processes (Kobus et al., 2001), (2) influence continuing investment even when costs are higher than profits (Staw, 1976; Staw & Ross, 1987), and (3) be moderated by individual differences in anxiety tolerance (Han et al., 2021). While some authors contend that ambiguity is the primary factor that exacerbates the SCF (Bossaerts et al., 2010; Haita-Falah, 2017), others assert uncertainty is a small contributor. O'Brien and Folta, 2009, contend that only when investments are substantial and the outcome is uncertain do sunk costs influence decisions. Lam and Yoon, 2021, argue that individuals with a high level of anxiety may continue to make comfortable decisions, and are particularly vulnerable to the SCF, rather than look for alternative possibilities to reduce the distress associated with options with ambiguous outcomes. Moreover, other research points to differences in cognitive ability that translate to a higher susceptibility

(Ronayne et al., 2021).

Project completion refers to the proportion of a project completed compared to that left to do. The closer a project is to completion, the less likely decision-makers will leave it (Arkes & Blumer, 1985; Garland, 1990; Garland & Conlon, 1998). Harvey and Victoravich, 2009, and Moon, 2001 argue that the SCF and project completion exert confounding effects on decisions, where both contribute to rejecting an alternate course of action, but for opposing reasons. That is, the SCF focuses on historical costs while project completion is a future-based outlook where a project that is further along is viewed with a higher certainty of success.

3. Experiment

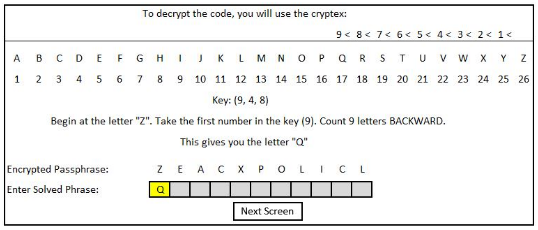
To test factors of influence on heuristics, and examine and measure the context in which the resultant biased decision-making occurs, we developed a new experimental platform: an abstracted cyber scenario we named CYPHER.

We posed the following research questions to examine the influence of uncertainty and project completion, to leverage the effects of the SCF.

- Research Question 1. Does uncertainty effect the sunk cost fallacy?
- Research Question 2. Does the project completion level effect the sunk cost fallacy?
- Research Question 3: Does the combination of uncertainty and project completion level effect the sunk cost fallacy?

3.1. Methods

Participants assumed the role of a cyber-defender to decrypt (i.e., solve) cipher text passphrases to catch a cyber attacker. This task was cognitively effortful, equivalent to the effort required in operational defensive activities. Participants had a limited number of resources (time, in-game currency) to use for decrypting the passphrases at a cost of one coin per correct entry, and two per error. The passphrases varied in length from 5 to 11 characters. In the training phase, participants learned how to use the alphanumeric table and key (fig. 1). The practice phase gave participants the opportunity to decrypt one passphrase without penalty (fig. 2). Finally, in the performance phase, participants decrypted seven passphrases on one of two data servers in two, 15-minute trials.



To decrypt the code, you will use the cryptex:

9 < 8 < 7 < 6 < 5 < 4 < 3 < 2 < 1 <

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26

Key: (9, 4, 8)

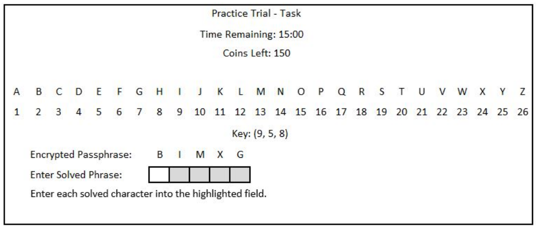
Begin at the letter "Z". Take the first number in the key (9). Count 9 letters BACKWARD.

This gives you the letter "Q"

Encrypted Passphrase: Z E A C X P O L I C L

Enter Solved Phrase: Q

Figure 1. Training Phase. Gronsfeld Cipher.



Practice Trial - Task

Time Remaining: 15:00

Coins Left: 150

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26

Key: (9, 5, 8)

Encrypted Passphrase: B I M X G

Enter Solved Phrase:

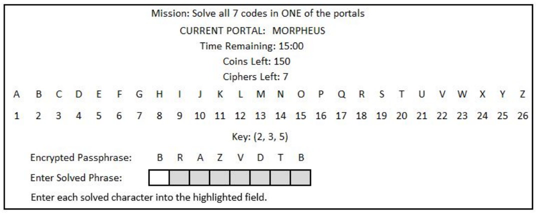
Enter each solved character into the highlighted field.

Figure 2. Practice Phase. Participants decrypted 1 cipher text passphrase containing 5 characters.

The task interface displayed resources (time and in-game currency), and number of passphrases remaining to be decrypted (fig. 3). Beginning with a bank of 150 coins, resources were refreshed between trials. The trial timer was paused for feedback and interactive messages. Following their successful decryption of a few passphrases, participants made the choice to switch to an alternate server or remain on the starting server. An interactive pop-up message contained information to assist participants in making this decision. Participants who chose to stay on the starting server, rather than switching to the alternate, indicated effects of the SCF. Qualitative surveys to enhance the quantitative interpretability were administered between the trials and following the completion of the performance phase.

3.2. Variables

The first variable of interest was the degree of certainty about the information describing the character length in each cipher text passphrase, on each server. Certainty was manipulated by presenting a pop-up message with credible and specific information



Mission: Solve all 7 codes in ONE of the portals

CURRENT PORTAL: MORPHEUS

Time Remaining: 15:00

Coins Left: 150

Ciphers Left: 7

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26

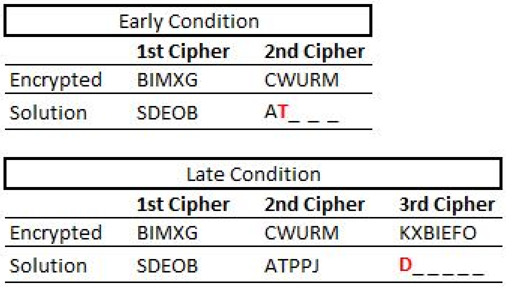
Key: (2, 3, 5)

Encrypted Passphrase: B R A Z V D T B

Enter Solved Phrase:

Enter each solved character into the highlighted field.

Figure 3. Performance Phase. Participants decrypted 7 cipher text passphrases on one of two servers.



Early Condition		
	1st Cipher	2nd Cipher
Encrypted	BIMXG	CWURM
Solution	SDEOB	AT_ _ _

Late Condition			
	1st Cipher	2nd Cipher	3rd Cipher
Encrypted	BIMXG	CWURM	KXBIEFO
Solution	SDEOB	ATPPJ	D_____

Figure 4. Early and Late presentation of the pop-up message.

whereas the uncertain information was of questionable credibility and unspecific. Using the information provided, participants were to calculate the best option for conserving resources. Once selected, respondents could not reconsider their decision.

The second variable of interest was project completion, or the ratio of complete to remaining work. Project completion was manipulated by the interval at which the pop-up message was presented, either earlier in the task, or later when more work had been done. The early condition received the message following the entry of the 2nd character in the 2nd passphrase (fig. 4). The late condition received the message following the entry of the 1st character in the 3rd passphrase.

The CYPHER experiment employed a 2x2 factorial design with quasi-random assignment to one of four conditions based upon (1) the certain or uncertain pop-up message descriptions of passphrase differences between the starting or alternate server and (2) the interval at which the project completion pop-up message was presented (early or later in the trial). The four conditions were: (1) Early-Certain, (2) Early-Uncertain, (3) Late-Certain, and (4) Late-Uncertain.

3.3. Qualitative Surveys

Qualitative data enhance quantitative interpretability by providing insight into hidden decision-making processes only observable as measurable behavior. Although the limitations of self-report must be considered, participant reflections on how and why decisions were made were essential for a more comprehensive understanding of participants' mental processes when completing tasks. Demographic information (e.g., age, education level, gender identity) and the following surveys were administered throughout the experiments. These data were interpreted to reflect the individual differences in participants' decision-making behavior alongside measures of the sunk cost fallacy.

3.3.1. Inter-trial Survey. At the end of trial 1, participants responded to three multiple-choice questions to gauge engagement and motivation. These questions were: *(1) "Did you enjoy this task?" (2) "How difficult would you say this task is?" and (3) "Why did you stay/switch portals?"*

3.3.2. End of Task Surveys. At the conclusion of the performance phase, participants responded to the following surveys.

- Task Questionnaire. The stimuli evaluated participants' intent and motivation, reasoning, and general risk-taking behavior (e.g., *"If you ever chose to switch to the other server during this experiment, how/why did you make that decision?")*
- Intolerance to Uncertainty Scale - Short Form 12 (IUS-12). This scale assesses two factors, (1) cognitive anxiety (prospective anxiety), and (2) behavioral anxiety (inhibitory anxiety) (e.g., *"When its time to act, uncertainty paralyzes me.")* (Carleton et al., 2012). These results would be used to determine significant differences in general uncertainty tolerance between groups that might influence decision-making.

3.4. Hypotheses

The following hypotheses guided our experiment:

- Hypothesis 1. The SCF occurs more in uncertain conditions. **(Supported in trial 2)**
  - Participants in the uncertain condition will switch less from the starting server to the alternate server than those in the certain condition.
- Hypothesis 2. The SCF occurs more when a project is closer to completion. **(Supported in trial 1)**
  - Participants in the late condition will switch less to the alternate server than those in the early condition.
- Hypothesis 3. The SCF occurs more in uncertain conditions when a project is closer to completion. (**Partially supported in trial 2)**
  - Participants in the Late-Uncertain condition will switch less to the alternate server than those in the Late-Certain and Early conditions.

3.5. Analysis

This section contains experiment recruitment, data collection, preparation and analysis.

Data collection for the experiment ran from July 14 through August 4, 2021. This research was conducted with the approval of the Institutional Review Board at Arizona State University (Study 00010523), and the Department of the Defense. Recruitment took place via Amazon MTurk, a crowd sourcing platform for human science research. Qualifications included English speaking, a Human Intelligence Task (HIT) approval rate of greater or equal to 95%, more than 50 completed HITs, and worker physical location limited to within the United States. A total of 6 HITs were posted, N = 875, which, after the data was prepared, n = 388. Participants by gender were Male, N = 243; Females, N = 144; Other, N = 1; and Age, M = 36.96; SD, 10.31; Min = 20, Max = 77. Due to the non-normality of our quantitative data, in each trial Chi-square tests of independence were performed to examine the relationship of uncertainty and project completion to staying on the starting server. Qualitative data were prepared and analyzed according to precedence from the originating studies.

3.5.1. Passphrase Description Task Data. Data more than 3 standard deviations from the mean, and for the following three criteria were removed prior to analysis:

Removal Criterion 1. Participants who did not complete the task as instructed (e.g., “key mashing). For example, these participants rapidly entered incorrect characters until resources were depleted (e.g., zero coins remained).

Removal Criterion 2. Participants who did not appropriately respond to survey questions. For example, the response was unrelated to the prompt, or the prompt response was copied from an outside source (e.g., unrelated web page) and pasted into the response field.

Removal Criterion 3. Identical, detailed survey responses from multiple participants that suggested the experiment details were shared amongst multiple participants, or that some participants had multiple MTurk accounts.

3.5.2. Qualitative Data. The Inter-trial Survey data were averaged across all participants. The Task Questionnaire data were summarized and grouped into general themes based upon key words and concepts such as explanations for staying on the starting server compared to switching to the alternative, motivation for study participation, and general risk taking attitudes (e.g., gambling and hobbies). The response frequency and Likert value for each question in the IUS-12 was averaged to provide an overall 12-question mean score.

3.6. Results

This sections provides the results for the performance phase (i.e., passphrase decryption task) and the qualitative surveys.

3.6.1. Passphrase Decryption Task. Overall, irrespective of condition, 159 participants chose to stay on the starting server in both trials compared to 118 who always switched. A paired samples t-test demonstrated a significant difference between those who always stayed (M = 14.45, SD = 4.46) and those who always switched (M = 10.73, SD = 2.83), t(10) = 2.17, *p* = .03.

Support for the hypotheses varied between trials. In trial 1, H1 was not supported, as the proportion of switching behavior did not significantly differ between the Certain and Uncertain conditions. H2 was supported as participants in the Late condition (62%) were significantly less likely to switch to the new server than the Early condition, χ^2 (1, N = 388) = 4.623, *p* = 0.03. This result suggests the project completion manipulation elicited the SCF. The results of the combined magnification effects of certainty and project completion were significant in trial 1 but in the opposite direction than hypothesized. That is, contradictory to H3, participants in the Late-*Uncertain* condition (60%) were more likely to switch compared to those in the Early-Uncertain condition, χ^2 (1, N = 388) = 3.80, *p* = .05.

In trial 2, in support of H1 the proportion of switching behavior was significantly less in the Uncertain condition (60%) compared to the Certain condition, χ^2 (1, N = 388) = 6.09, *p* = 0.01. These results suggest uncertainty elicited the SCF. H2 was not supported, as the difference between Early and Late conditions did not significantly differ. The proportion of difference between the combined effects again contradicted H3, but where those in the Late-*Certain* condition (61%) were more likely to switch than the Early-Uncertain condition χ^2 (1, N = 388) = 5.02, *p* = .03.

3.6.2. Qualitative Surveys. An overview of the findings from the qualitative surveys are presented in this section. Future work should involve a more fine-grained analysis and a comparison between groups.

Inter-Trial Survey. In summary, most participants reported enjoying the experiment (79%), found that solving the passphrases was moderately easy (57%), and switched to the alternate server because they believed it would be faster (61%). Switching portals as a strategy to save time aligns with what is generally believed about the MTurk subject pool: workers are motivated to complete tasks in as little time as possible. We speculate, for some participants, regardless of whether passphrase decryption required more resources or not, participants tended toward the strategy they believed to be faster – even when this belief was false.

Task Questionnaire. The purpose of the free response questions was to gather thoughts about the decision-making processes when deciding to stay on the starting server or switch to the alternate. In general, most participants reported making decisions related to the manipulated variables, but others' decisions were based upon unrelated factors such as curiosity, time, and perceived effort. For example, some participants reported they preferred to *"finish what I start"* or they *"trusted the advice provided"* in the pop-up message and chose the alternate server. Others stated they *"were curious what was on the other server."*

Intolerance to Uncertainty Scale. When comparing groups, a higher score would suggest that participants who had a lower tolerance for uncertainty might be more inclined to stay on the starting portal if they perceived switching to be too risky, regardless of the condition. We did not find a significant difference between conditions, F(3, 47) = .64, *p* = .60.

While no pre-test was given to determine participants' intolerance to uncertainty prior to completing the experiment, we take these results to suggest that following the performance phase, all reported similar responses indicative of no relevant difference between groups. Interestingly, compared to the normative population for the scale, the mean scores in our sample demonstrated traits associated with panic disorder. This indicates a higher-than-average level of anxiety, particularly in situations of uncertainty (Carleton et al., 2012).

To make sense of this result, we consider that these data were collected during the COVID-19 pandemic. Recent research demonstrates COVID-19 produced societal and economic uncertainty and reportedly increased anxiety in every-day responsibilities. During the pandemic years, anxiety levels in the United States increased 3-fold (Santabárbara et al., 2021). We suggest our participants' high IUS-12 scores were situationally related, rather than inherent qualities of MTurk workers.

4. Discussion

Overall, the crosstabulation demonstrates that irrespective of condition, there were significantly more participants who stayed on the starting server compared to those that switched to the alternative. This means a significant proportion unnecessarily wasted resources to achieve their prescribed goal. Furthermore, we found support for the individual contribution of uncertainty or project completion, but the combined effects of uncertainty and project completion were mixed.

- Hypothesis 1: In trial 1, we found evidence that higher project completion related to persisting on the starting server, even though doing so cost more resources than switching to the alternate.

- Hypothesis 2: In trial 2, we found evidence that uncertainty about the alternate server related to persisting on the starting server, which also wasted resources.

- Hypothesis 3: In trial 2, participants who received the pop-up message with *uncertain* information at the *earlier* interval wasted more resources than those who received the pop-up message with *certain* information at the *later* interval. We speculate the *content* of the certainty message exhibited a stronger influence than the *timing* of when it was displayed

This introductory study provides insight into participants' decision-making processes and behavior in a cyber task. We speculate that these results show the effects of SCF are present and measurable in cyber environments. However, we have not yet captured some of the environmental factors like effort avoidance, contextual elements, and experience with cyber tasks. We put forward this work as a starting point to validate the foundational theory in behavioral economics, and to confirm the SCF is present and measurable in a cyber task environment. Other researchers (Plonsky et al., 2019) and (Bourgin et al., 2019) confirm the complexity of measuring decision-making heuristics and biases in that "theoretical explanations for deviations" from behavioral economics theory "are often contradictory, making it difficult to come up with a single framework that explains the plethora of empirically observed deviations." Replication and consideration for additional factors is required to determine if the effects seen across the participants in this experiment extend to those trained to complete cyber tasks.

With these data collected from human science research, we offer the means to establish a human behavior baseline for pre-training computational models to advance automated cyber defenses. Using ACE as a theoretical foundation to understand attacker cognitive processes, cyber defensive tactics can specify the techniques for exploiting vulnerabilities to generate strategically relevant effects. We posit that computational cognitive models characterize behavioral data in terms of latent, underlying variables associated with hypothesized cognitive processes. Such models quantify abstract cognitive concepts and permit the integration of cognitive theory and human behavioral data within a statistical framework. These frameworks can provide a foundation for technological solutions like decision analytics, and AI/ML capabilities that are a scalable, automated, and predictive. Furthermore, in critical mission organizations that must adapt quickly to changing situations, robust automation grounded in reliable and validated human science data is a necessary step to advance the efficacy of defensive cybersecurity.

5. Lessons Learned

It is challenging to create a novel experimental platform to investigate the antecedents such as uncertainty and project completion, that give rise to heuristics and the resultant biases in decision-making in a cyber-specific domain. In general, this experiment demonstrated mixed results. Beyond the complexities and probable myriad of factors in human decision-making, we offer the following lessons learned that may be useful for future research design.

Limited Trials. We recognize that learning effects and novelty may contribute to inconsistent results in laboratory experiments. Although we provided a training and practice phase, these experiences may have been too brief to overcome the novelty of the game. Furthermore, the performance phase had only two trial with one opportunity to measure the decision to stay on the starting portal or switch to the alternative. Future research may reduce these effects and capture more representative behavior by increasing the practice phase and number of trials.

Non-Salient Variables. According to survey results, we speculate that for certain participants, the experimental variables were less salient than the game story line, while others might have lacked sufficient comprehension of the text, most critically in the pop-up message prompts. For example, the narrative text may have distracted participants from the embedded task instructions. Based on self-report, found evidence of this as some participants reported confusion about the calculations required to make the best decision between staying with their current set of cipher text or switching to the alternative server. A more careful inspection of the provided information would have revealed the resource conservation by switching (35 characters) compared to staying (approximately 45 characters). According to Loepp and Kelly, 2020, based upon their sample of 600 participants, the mean education level of regular MTurk U.S. workers is a 4-year college degree. Hence, our assumption that the mental calculations required by the switch message were appropriate. However, we surmise the presentation of the critical information caused unintended effects in some participants (Anson, 2018).

Comprehension of text. Based on self-report, it is possible the presentation of large bodies of text and font size limitations was an issue. We used a sans serif size 12 font. Yet, the resolution and monitor size used to complete the HIT may have varied. We do not know how this may have affected task performance, but speculate that some participants may have experienced reduced readability.

Pop-Up Message Presentation. The difference in the number of characters between the Early and Late presentation was five. We believe the timing of Late condition presentation of the pop-up message was not different enough from the Early presentation and thus, not a good comparison. In future experiments, we suggest manipulations of decision-making intervals should be carefully evaluated to consider both the the real and perceptual differences.

Participant Pool. Amazon MTurk is primarily a micro-tasking platform. Workers often select tasks that can be completed quickly. It is possible the complexity and effort to decrypt passphrases presented enough novelty to attract participants, but required more effort than expected. Although 875 participants completed the HIT, only 388 provided quality data for analysis.

Data Collection. We began data collection while requiring the "Masters" Certification, which is awarded to Mturk workers by Amazon via a threshold of performance over multiple HITs. The data collected was clean, participants followed instructions and completed the task as required. However, the rate at which we could collect was very slow. We made the decision to remove the requirement, restart data collection and perform the necessary data cleaning on the back end. While this exponentially increased our rate of collection, the data received was also more variable, and resulted in data removal thereby requiring a much larger number of participants. The issues surrounding methods for participant recruitment and additional performance-based methods to incentivize participants should be explored in future work.

Platform Design A weakness related to participant data quality was found in the inherent limitations of experiment crowdsourcing. Without direct control, respondents could share cipher text solutions online, write an application to rapidly decrypt the cipher text, or use other tools to greatly reduce the cognitive effort required in the task. However, these risks were hopefully ameliorated by rigorous quality evaluation across all tasks and qualitative survey responses. MTurk HITs containing inconsistencies, data duplication and survey responses that indicated dishonest responses were rejected and/or removed from data collection records. A major limitation to this method is the extensive amount of data curation that must occur in a limited time window (e.g., 24 hours from HIT completion) for participant compensation. If poor data was not rejected in a timely manner, payment was rendered even though the data was unusable. The administrative burden had direct implications for experiment costs including personnel, time, and monetary resources.

6. Conclusion

As technological solutions for cybersecurity and defense turn toward automated techniques, cognitive models based on human science research data provide a valid foundation for the ways humans actually think and behave. This is a critical component in supporting effective human interaction with tools and systems. These data also advance technology development by increasing the learning efficiency in machine learning models (Schürmann & Beckerle, 2020). Here, we introduce the CYPHER experiment: an initial approach to understand decision-making under the SCF, influenced by factors of uncertainty and project completion.

We found that overall, a significant number of participants demonstrated the effects of the SCF by wasting resources to reach their goal. The results validate (1) the foundational behavioral economics theory is applicable in a cyber task and (2) using Adversarial Cognitive Engineering methodology to exploit mental process affects decision-making that give rise to measurable bias. Importantly, the results suggest this methodology may be used to thwart attacker goals and strengthen cyber defense strategies.

We learned much about experiment design for investigating heuristics and biases. We selected two factors from the decades of research on the SCF and found the results were not as clear cut as we hoped. We also learned the benefits and limitations of crowdsourcing platforms for data collection. Initially, our research was developed for in-person experimentation, but the necessary restrictions of the Covid 19 pandemic required we re-engineer our platform for an online crowdsourcing platform. In our experience, we had to choose between a capped number of the highest qualified participants and those who were more readily available but with more qualification variance. Importantly, the findings provide a viable avenue for real and practical application to future technology development and support advancements in cyber defense by affecting adversarial goals and outcomes.

References

Aharoni, Y., Tihanyi, L., & Connelly, B. L. (2011). Managerial decision-making in international business: A forty-five-year retrospective. *Journal of World Business*, *46*(2), 135–142.

Anson, I. G. (2018). Taking the time? explaining effortful participation among low-cost online survey participants. *Research & Politics*, *5*(3), 2053168018785483.

Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological bulletin*, *110*(3), 486–498.

Arkes, H. R., & Blumer, C. (1985). The psychology of sunk cost. *Organizational behavior and human decision processes*, *35*(1), 124–140.

Augier, M., & Teece, D. J. (2021). *The palgrave encyclopedia of strategic management*. Palgrave Macmillan.

Bossaerts, P., Ghirardato, P., Guarnaschelli, S., & Zame, W. R. (2010). Ambiguity in asset markets: Theory and experiment. *The Review of Financial Studies*, *23*(4), 1325–1359.

Bourgin, D. D., Peterson, J. C., Reichman, D., Russell, S. J., & Griffiths, T. L. (2019). Cognitive model priors for predicting human decisions. *International conference on machine learning*, 5133–5141.

Carleton, R. N., Mulvogue, M. K., Thibodeau, M. A., McCabe, R. E., Antony, M. M., & Asmundson, G. J. (2012). Increasingly certain about uncertainty: Intolerance of uncertainty across anxiety and depression. *Journal of anxiety disorders*, *26*(3), 468–479.

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American psychologist*, *49*(8), 725–747.

Ferguson-Walter, K. J., Major, M. M., Johnson, C. K., & Muhleman, D. H. (2021). Examining the efficacy of decoy-based and psychological cyber deception. *30th USENIX Security Symposium*, 1127–1144.

Friedman, D., Pommerenke, K., Lukose, R., Milam, G., & Huberman, B. A. (2007). Searching for the sunk cost fallacy. *Experimental Economics*, *10*, 79–104.

Garland, H. (1990). Throwing good money after bad: The effect of sunk costs on the decision to escalate commitment to an ongoing project. *Journal of Applied Psychology*, *75*(6), 728–731.

Garland, H., & Conlon, D. E. (1998). Too close to quit: The role of project completion in maintaining commitment 1. *Journal of Applied Social Psychology*, *28*(22), 2025–2048.

Gomez, M. A., & Whyte, C. (2022). Unpacking strategic behavior in cyberspace: A schema-driven approach. *Journal of Cybersecurity*, *8*(1), tyac005.

Gutzwiller, R., Ferguson-Walter, K., Fugate, S., & Rogers, A. (2018). 'Oh, Look, A butterfly!' A framework for distracting attackers to improve cyber defense. *Human Factors and Ergonomics Society (HFES)*.

Haita-Falah, C. (2017). Sunk-cost fallacy and cognitive ability in individual decision-making. *Journal of Economic Psychology*, *58*, 44–59.

Han, P. K., Strout, T. D., Gutheil, C., Germann, C., King, B., Ofstad, E., Gulbrandsen, P., & Trowbridge, R. (2021). How physicians manage medical uncertainty: A qualitative study and conceptual taxonomy. *Medical decision making*, *41*(3), 275–291.

Harvey, P., & Victoravich, L. M. (2009). The influence of forward-looking antecedents, uncertainty, and anticipatory emotions on project escalation. *Decision Sciences*, *40*(4), 759–782.

Johnson, C. K. (2022). *Decision-making biases in cybersecurity: Measuring the impact of the sunk cost fallacy to disrupt attacker behavior* (Doctoral dissertation). Arizona State University.

Johnson, C. K., Gutzwiller, R., Ferguson-Walter, K., & Fugate, S. (2020). A cyber-relevant table of decision making biases and their definitions [https://www.researchgate.net/publication/344106644_A_Cyber-Relevant_Table_of_Decision_Making_Biases_and_their_Definitions]. https://doi.org/10.13140/RG.2.2.14891.87846

Kahneman, D. (2011). *Thinking, fast and slow.* Farrar, Straus and Giroux.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American psychologist*, *64*(6), 515–526.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American psychologist*, *39*(4), 341–356.

Klein, G. A., Calderwood, R., & Clinton-Cirocco, A. (1986). Rapid decision making on the fire ground. *Proceedings of the human factors society annual meeting*, *30*(6), 576–580.

Knight, F. H. (1921). *Risk, uncertainty and profit* (Vol. 31). Houghton Mifflin.

Knott, B. A., Mancuso, V. F., Bennett, K., Finomore, V., McNeese, M., McKneely, J. A., & Beecher, M. (2013). Human factors in cyber warfare: Alternative perspectives. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 399–403.

Kobus, D. A., Proctor, S., & Holste, S. (2001). Effects of experience and uncertainty during dynamic decision making. *International Journal of Industrial Ergonomics*, *28*(5), 275–290.

Lam, J. C., & Yoon, K. L. (2021). Why change now? cognitive reappraisal moderates the relation between anxiety and resistance to sunk cost. *Journal of Psychopathology and Behavioral Assessment*, *43*, 314–319.

Loepp, E., & Kelly, J. T. (2020). Distinction without a difference? an assessment of mturk worker types. *Research & Politics*, *7*(1), 2053168019901185.

Mancuso, V. F., Funke, G. J., Strang, A. J., & Eckold, M. B. (2015). Capturing Performance in Cyber Human Supervisory Control. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *59*(1), 317–321. https://doi.org/10.1177/1541931215591066

Mancuso, V. F., Strang, A. J., Funke, G. J., & Finomore, V. S. (2014). Human factors of cyber attacks: A framework for human-centered research. *Proceedings of the human factors and ergonomics society annual meeting*, *58*(1), 437–441.

Moon, H. (2001). Looking forward and looking back: Integrating completion and sunk-cost effects within an escalation-of-commitment progress decision. *Journal of Applied psychology*, *86*(1), 104–113.

O'Brien, J., & Folta, T. (2009). Sunk costs, uncertainty and market exit: A real options perspective. *Industrial and Corporate Change*, *18*(5), 807–833.

Plonsky, O., Apel, R., Ert, E., Tennenholtz, M., Bourgin, D., Peterson, J. C., Reichman, D., Griffiths, T. L., Russell, S. J., Carter, E. C., et al. (2019). Predicting human decisions with behavioral theories and machine learning. *arXiv preprint arXiv:1904.06866*.

Ronayne, D., Sgroi, D., & Tuckwell, A. (2021). Evaluating the sunk cost effect. *Journal of Economic Behavior & Organization*, *186*, 318–327.

Santabárbara, J., Lasheras, I., Lipnicki, D. M., Bueno-Notivol, J., Pérez-Moreno, M., López-Antón, R., De la Cámara, C., Lobo, A., & Gracia-Garcıa, P. (2021). Prevalence of anxiety in the covid-19 pandemic: An updated meta-analysis of community-based studies. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *109*, 110207.

Schürmann, T., & Beckerle, P. (2020). Personalizing human-agent interaction through cognitive models. *Frontiers in Psychology*, *11*, 561510.

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, 99–118.

Slovic, P., Finucane, M., Peters, E., & MacGregor, D. G. (2002). Rational actors or rational fools: Implications of the affect heuristic for behavioral economics. *The journal of socio-economics*, *31*(4), 329–342.

Staw, B. M. (1976). Knee-deep in the big muddy: A study of escalating commitment to a chosen course of action. *Organizational behavior and human performance*, *16*(1), 27–44.

Staw, B. M., & Ross, J. (1987). Behavior in escalation situations: Antecedents, prototypes, and solutions. *Research in Organizational Behavior*, *9*, 39–78.

Steingartner, W., Galinec, D., & Kozina, A. (2021). Threat defense: Cyber deception approach and education for resilience in hybrid threats model. *Symmetry*, *13*(4), 597.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *science*, *211*(4481), 453–458.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, *5*, 297–323.