The Enterprise Strikes Back: Conceptualizing the HackBot - Reversing Social Engineering in the Cyber Defense Context

Michael J. Lundie
Applied Research Associates, Inc.
mlundie@ara.com

Mary P. Aiken
Capitol Technology University
mpaiken@captechu.edu

Adam Amos-Binks
Applied Research Associates, Inc.
aamosbinks@ara.com

Kira L. Lindke
Applied Research Associates, Inc.
klindke@ara.com

Diane M. Janosek
Capitol Technology University
janosek.diane@gmail.com

Abstract

Cyberattacks have become more complex and pervasive; associated costs are soaring; there is an urgent need for innovative solutions. Socially engineered attacks are escalating in scale, potency, and are increasing in frequency; defenses have not evolved and tactics currently deployed are passive, and arguably offer little deterrent value. Social engineering is rooted in psychology and mediated by technology, therefore, solutions must be informed by a transdisciplinary approach, with the cyber behavioral sciences taking a central role. Identifying and targeting cyberattacker psychological vulnerabilities by means of active cyber defense are under consideration. Automation and scale of response are key requirements, underscoring the need for and the utility of large language models (LLM), in terms of identifying context, scaling to attack type, and generating dialogue to engage the cyberattacker and effectively 'hack back.' Hence the present conceptualization of the "HackBot" - an automated strike back innovation, specifically devised to reverse socially engineered attacks in cyber defense contexts.

Keywords: cyberattack, large language models (LLM), cyberpsychology, cybersecurity, psychological vulnerability

1. Introduction

The threat of cyberattacks is ubiquitous and spans a broad range of targets threatening critical infrastructure, government institutions, and individuals. The consequences of cyberattacks have become more severe as the capabilities of attackers become more sophisticated, causing disruption to enterprise operations, threatening customer relations due to data breaches, diminishing contract revenue, devaluing trade names, and compounding losses of intellectual property. The rapid growth of the use of digital technologies and data-based solutions has increased the scale of cyber threats and brought new challenges, additionally, the effects of potential cyberattacks have also become more complex and spread more widely (Deloitte, 2023). The costs associated with data breaches and financial losses have prompted major realignment of resources toward mitigating risks. The Director of the Cybersecurity and Infrastructure Security Agency (CISA) recently stated that "cyber threats to the systems that control and operate the critical infrastructure we rely on every day are among our greatest challenges. As the destruction or corruption of these control systems could cause grave harm, ensuring their security and resilience must be a collective effort" (Cyber and Infrastructure Security Agency, 2022, para. 2), highlighting the need to tap into the innovation, expertise, and ingenuity of the industrial control systems community. The global average cost of a data breach went up 2.6% from $4.24M in 2021 to $4.35M in 2022, this cost is the highest recorded to date (IBM Security, 2022). Industry, government and academia alike are exploring all legal options to lower costs and mitigate harms associated with cyberattacks.

The threat landscape encompasses attack vectors of various sorts, many relying upon technical acumen to hack into vulnerable systems across sectors, systems and enterprises (Janosek, 2021). However, the vast majority of cyberattacks rely on exploiting human operators within organizations using deceptive attacks in order to carry out an offensive operation. Indeed, socially engineered attacks constitute 98% of all phishing and data breach operations (Proofpoint, 2023). Social engineering refers to the use of deception

URI: https://hdl.handle.net/10125/106496
978-0-9981331-7-1

by a cyberattacker, scammer, or fraudster to target an individual and manipulate them to gain access to a system, thereby jeopardizing the proprietary information or financial resources of the targeted victim or organization (Aiken et al., 2022; Mouton et al., 2016). Social engineering is implemented by leveraging the default human tendency to trust others. Manipulation of trust is deployed to acquire information assets. Notably, the amount of social engineering attacks and associated damage rises every year, yet defenses against social engineering have not evolved accordingly (Beckers et al., 2017).

1.1. Paradigm shift from passive to active cyberdefense

Arguably to date, defenses against cyberattacks have been characterized by what could be described as a 'passive defense posture' (Djekic, 2023; McLaughlin, 2011), which includes defense against social engineering tactics, encompassing multi-factor authentication; augmenting password strength; updating malware protection and education campaigns within organizations (Proofpoint, 2023). In the context of security breaches, network traffic is typically captured and monitored using PCAP (packet capture) and IDS (Intrusion Detection System) alerts. The problem is that these formats generate large volumes of data, thus presenting difficulties with managing, prioritizing and responding to alerts (Zaccaro et al., 2016).

Almost a decade ago Ahmad et al. (2012) argued that cybersecurity had mostly aligned with technical advancement considerations, rather than a psychosocial approach which could target vulnerabilities of attack vectors. McAlaney et al. (2016) noted that despite the intrinsic psychological nature of many cybersecurity attacks, research into the role of psychology in cybersecurity was very limited, adding that research into social engineering has been paradoxically mostly conducted from the discipline of computing rather than the behavioral sciences. This paradigm has emphasized the potential utility of emerging interdisciplinary fields such as cyberpsychology (Connolly et al., 2016) and forensic cyberpsychology. It has been argued that "the critical task for cyberpsychology as a discipline is to build up a body of established findings of how human beings experience technology, the critical task in forensic cyberpsychology is to focus on how criminal populations present in cyber environments" (Aiken & Mc Mahon, 2014, p. 82). An innovative Intelligence Advanced Research Projects Activity (IARPA) project titled Reimagining Security with Cyberpsychology-Informed Defenses (ReSCIND) has launched a landmark cyberpsychology-informed research initiative to study cyberattackers' psychological weaknesses and exploit them. Initiatives such as ReSCIND aim to "apply traditional cognitive behavioral science research — now mediated by cyberpsychological findings and learnings — and apply that to cybersecurity to improve defensive capabilities" (Groll, 2023). Furthermore, the recent 2023 Computational Cybersecurity in Compromised Environments conference focused on cyberpsychological aspects of foreign malign influence; innovation in AI and machine learning; generative AI and large language models.

1.2. Active cyber offensive operations

As stated, due to the high cost and disruption of cyberattacks, researchers are now assessing whether disruptive cognitive techniques aimed at the attacker's mental limits and biases could be applied (IARPA, 2023). Notably, a recent National Cyber Force (NCF) report outlined how the UK is taking a new approach to conducting offensive cyber operations with a focus on disrupting information environments. This strategy introduces what is described as a "doctrine of cognitive effect" (National Cyber Force, 2023, p. 15), which aims to counter adversarial behavior by exploiting reliance on digital technology. In doing so, offensive cyber operations can limit an adversary's ability to collect, distribute, and importantly, trust information (Fendorf & White, 2023). These new approaches may deliver effective interventions. However, the questions to be considered are as follows; firstly, could the application of disruptive cyberpsychological techniques be effective across a range of cyberattacker behaviors? Secondly, how could concerns be mitigated regarding violation of legal boundaries in terms of active offensive cyber operations?

Depending on the jurisdiction, rules differ on legality, but generally active cyber operations are restricted, especially when seen as an act of aggression not an act of defense (Schmitt, 2017). When cyber responses cross into the area of sabotage of a country's critical infrastructure, such as water or energy, or impact a sovereign's core capability, the cyber action will be deemed aggressive or hostile under current norms. Once attribution is confirmed, hostile cyber aggressions could then presumably be met with internationally accepted counterattacks, and these are not limited to non-kinetic cyber responses (Willett, 2023). Cyber aggressions that occur during wartime are responded to, generally in kind, by nation states. However, a cyberattacker is often not a nation state actor, at least not overtly, and the victim is often not a sovereign entity. This creates the legal conundrum as to how far along the spectrum of cyber defense and cyber offense can a private enterprise proceed, especially without certainty of attacker attribution. There is a need for further research on the limits or boundaries of cyber self-defense. Such legal and policy research should inform HackBot protocols to ensure an appropriate level of response. However, this subject area is very complex and indeed constitutes a research topic in its own right, encompassing: prevention of hacking back against benign users; machines being used in botnets; standardizing levels of retaliation to avoid excessive revenge hacking; along with a full review of the global legal parameters of active cyberdefense.

Broeders (2021, p. 1) notes that private sector Active Cyber Defence (ACD) "lies on the intersection" of domestic security and international security and is a "recurring subject, often under the more provocative flag of 'hack back'... Corporate self-help in cyberspace is a contentious issue." Industry may suffer as a result of cyberattacks by a range of threat actors (cybercriminals or state-sponsored actors). However, Broeders (2021) observes that the private sector can defend their networks but are not permitted to follow or retaliate beyond the perimeter of their own networks. This premise is grounded in legislation such as the American Computer Fraud and Abuse Act (CFAA) (1984). In the U.S., there exists permissible cyber offensive operations, however they are generally limited to U.S. agencies in support of approved law enforcement and special intelligence operations (Willett, 2023). Non-compliance can be enforced with criminal penalties as codified in U.S. Code (Authorities Concerning Military Cyber Operation, 2018). Outside of the U.S., countries differ on limitations. In the UK, active offensive cyber operations with kinetic effects cannot be undertaken by private entities (Sciacovelli, 2022). Cyber operations may not cause direct physical harm to humans and property, such as blowing up a bridge through a cyberintrusion of drawbridge operator's controls, especially when seen as unprovoked aggression (Sciacovelli, 2022). Accordingly, the appropriate range of automated HackBot responses will need to be refined to ensure legal and ethical compliance across the globe, and specifically psychological ethical implications.

1.3. Active defense and hacking

In terms of active defense and 'hacking back' it has been argued that merely aspiring to construct impregnable network defenses is not sufficient to ward off competent cyberattackers and the most capable advanced persistent threats (APTs) (Berinato, 2018). NATO defines active defense as: "a proactive measure for detecting or obtaining information as to a cyber intrusion, cyber attack, or impending cyber operation or for determining the origin of an operation that involves launching a preemptive, preventive, or cyber counter-operation against the source" (Berinato, 2018). Operational Psychology has been defined as the "specialty within the field of psychology that applies behavioral science principles to enable key decision makers to more effectively understand, develop, target, and/or influence an individual, group or organization to accomplish tactical, operational, or strategic objectives within the domain of national security or national defense" (Staal & Stephenson, 2013, p. 97) and is known in defense terms as psychological operations (PSYOPs). Operational cyberpsychology is an emerging field which "supports missions intended to project power in and through cyberspace…by leveraging and applying expertise in mental processes and behavior in the context of interaction amongst humans and machines" (Spitaletta, 2021, p. 3), in terms of active defense and deploying cyberpsychological 'hack backs' this could now perhaps be conceptualized as "CyberPSYOPs."

1.4. Cognitive disruptive operations

In today's global cybercriminal enterprise, there exist human vulnerabilities which can be targeted and exploited to disrupt adversarial intentions and chances of success, for example defensive cyber deception (Ferguson-Walter et al., 2021). The goal therefore is to achieve effective disruptive defensive operations without breaching the threshold of active cyber offensive operations. Thus, it is imperative to establish a range of methodologies to diminish attackers' potential for success. In what could be described as a form of cognitive disruptive operations, options are being explored to undermine and exploit cyber attacker cognitive biases (IARPA, 2023).

Cyberattacks have become more sophisticated (Deloitte, 2023); associated costs are soaring (IBM Security, 2022); there is an urgent need for innovative solutions (Cyber and Infrastructure Security Agency, 2022); socially engineered attacks are pervasive (Bergler et al., 2021); and defense tactics are passive (National Cyber Force, 2023). Social engineering is rooted in psychology and mediated by technology, therefore, solutions must be informed by the cyber behavioral sciences (Aiken & Mc Mahon, 2014; Martineau et al., 2023; National Cyber Force, 2023). Active defense by means of identifying and targeting attacker psychological vulnerabilities are being explored (Berinato, 2018; Fendorf & White, 2023; Groll, 2023; IARPA, 2023; National Cyber Force, 2023) automation and scale are key concerns and the ability to hack back is gaining traction (Berinato, 2018; Groll, 2023; IARPA, 2023; National Cyber Force, 2023), hence the present conceptualization of the "HackBot" - a theoretical automated strike back innovation, specifically devised to reverse socially engineered attacks in cyber defense contexts. The following section outlines context, scale, dialogue, and experimental processes required to develop future HackBot capabilities.

2. HackBot: a prototype counter-social engineering cyber defense system

The HackBot's task is to generate text that adapts to a socially engineering attack. Challenges include identifying the context of a specific attack type, scaling to a wide range of attacks, and generating the hallmark/stereotypical dialogue of the attacker's target. One approach is to leverage pre-trained large language models (LLM) and fine-tune with incident reports of social engineering attacks. LLMs (Cer et al., 2018; Ouyang et al., 2022; Tenney et al., 2019) are ideally suited to overcome the above challenges due to their wide availability, low need for examples of downstream tasks, and nimbleness to adapt to new contexts.

Context: A key property of LLMs is representing language as a finite, high-dimension vector of numbers that is referred to as an embedding, for example by using the attention mechanism (Vaswani et al., 2017). Embeddings represent the semantics of natural language text and facilitate using mathematical comparisons. In this case, the opening language used by a social engineering attack script is easily compared to historical examples of attacks (e.g. incident reports). The numerical representation simplifies comparisons between historical and current language, accounting for variations in attack script variables (e.g. company name) and synonyms without explicitly having to hand-label historical examples. The result is the capability to identify a social engineering attack in its early stages so as to begin engaging in a context that is desirable (i.e. high likelihood of exploit) to an attacker.

Scale: LLMs are developed using many historical examples of natural language and as a result are very general. This general capability can be refined to represent the breadth of social engineering attack types. Fine-tuning is the practice of modifying the semantics of a LLM's embedding (Houlsby et al., 2019) to more appropriately weight the representation, in terms of the HackBot this is attack text. In technical terms, this requires examples of the natural language used in a social engineering attack to use as ground truth in a supervised learning task to fine-tune. When there are few examples of attack types, the general nature of LLMs can be leveraged to identify types of social engineering cyberattacks with very few examples, known as k-shot learning where k is the number of examples. Therefore, LLMs hold a promising avenue of investigation for scaling up to identify social engineering attacks.

Dialogue: LLMs can be used as a specific type of machine learning model known as sequence-to-sequence models (Wolf et al., 2020) Input is natural language, a sequence of words or sentences, that is transformed into an LLM's embedded format from which the output sequence is generated (also natural language text). In the present case, the output sequence must both engender trust in an attacker by feigning hallmarks of a target with high-likelihood exploit and elicit information useful to exploit the attacker - in other words, to 'hack back.' This is an ambitious goal, and with little margin of error. Careful consideration about the cyberattacker's mental state, such as psychological vulnerabilities, must be made in the output sequence. At the slightest chance of suspicion, an attacker might abort their attack. Although this is a win for the defender, an opportunity to consume the attacker's resources is lost. The HackBot must simultaneously be conducting, perhaps with human supervision, an exploit operation against the attacker and incorporating knowledge from the attack lifecycle into the output sequence of text.

LLMs have desirable qualities that facilitate a focus on the automated and flexible capability to 'hack back' by exploiting cognitive vulnerabilities without the burden of troves of historical data. In the short term, a HackBot can waste an attacker's resources on a decoy and in the long-term exhaust, disrupt and foil their capabilities by exploiting their own network infrastructure. To accomplish these two objectives while engaging in the scenarios (see Table 1) requires LLMs to integrate Theory-of-mind capability (the ability to reason about and infer another agent's mental state). This area is yielding interesting preliminary results (Shapira et al., 2023).

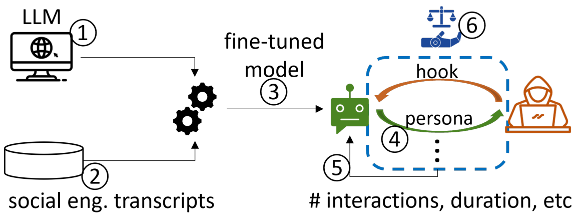

Figure 1. HackBot experimental design

Experimental and Design Process: ① (see Figure 1) The HackBot as conceptualized initiates with a pre-trained large language model (LLM), providing the general reasoning capability needed to generate natural language that is engaging to an attacker. ② This capability is supplemented with incident report transcripts of social engineering attacks, in order to ensure the HackBot can maintain a natural language conversation with the attacker in the appropriate attack context. There are several examples of LLMs being one-shot learners (e.g. a single example providing the context needed) and obtaining context from fine-tuning (e.g. only a few examples needed when compared to the LLM data). ③ While there is cause to be optimistic about the small amount of data needed to obtain high performance it will not be known from the outset how many historical incident report examples will be necessary to fine-tune an LLM to engage attackers (this will be bench-marked during research and design phases). ④ Once engaged the HackBot both infers context from the attacker's natural language text (e.g. attacker's hook, motivation for contacting, info requested, etc.), adjusting generated responses to gather more information and ensure the attacker stays engaged. ⑤ Readily available measures of effectiveness will be used to adjust the HackBot model using a method called Reinforcement Learning with Human Feedback (RLHF), including length of interaction and number/length of messages elicited from the attacker, both of which are proxies for wasting an attacker's resources. More longitudinal measures such as a reduction in total or success of attacks would be valuable (Measurement and exploitation metrics will be defined during the design and testing process). ⑥ An automated governance structure will be put in place (informed by extensive legal review), ensuring ethical, legal, and compliance considerations are integrated into HackBot protocols. Future research design and testing process will address the following: generation of a model that can predict a cyberattacker's campaign; types of information collected to determine cyberattacker goals; data needed for accurate determination; measures of 'hack-back' effectiveness; permissions or tools required to respond effectively and choice of response to ensure maximum effectiveness.

3. Exploiting psychological vulnerabilities of cyberattackers using HackBot

This position paper advocates a paradigm shift from passive to active forms of defense against cyberattack using the HackBot, that is, a specialized NLP application designed to counter social-engineering cyberattacks. As discussed, cybercriminals exploit psychological vulnerabilities to carry out socially engineered attacks (Montañez et al., 2020). Logically, the reverse may also apply. Human cyberattackers have a wide range of psychological vulnerabilities and therefore, hypothetically, these could be identified and targeted in terms of reversing the process or, in the context of this position paper, 'hacking back.' Arguably, the automation of such defense tactics by means of the development of a HackBot would allow for active defensive operations at scale. Phillips et al. (2022) summarize an extensive range of cybercrimes, of which cyberattacks are a subset, ranging from cyber-dependent to cyber-enabled crimes, and from cyber fraud to cyber stalking that could be detected and repelled by automated defenses deployed by a HackBot.

The cybersecurity domain integrates psychological and human elements. Understanding the human element of offense is critical to informing good defense (Zaccaro et al., 2016). Embracing this imperative of informing cybersecurity with behavioral sciences, IARPA has identified loss aversion and the representativeness heuristic as cognitive biases that may be targeted by cyberpsychology-informed cyber defenses (IARPA, 2023). Active offensive capabilities potentially deployable by the HackBot push beyond cognitive biases to explore a broader range of psychological vulnerabilities and draw on theoretical and empirical research from domains across the behavioral sciences. The hacker mentality for example is often characterized by high levels of anxiety, paranoia, and risk taking (Aiken et al., 2023a; Aiken et al., 2016; Martineau et al., 2023). The Hackbot will be designed to identify and neutralize cybersecurity attacks informed by cyberattacker psychological vulnerabilities.

3.1. Leveraging psychological vulnerabilities of cyberattackers

The HackBot may be used in conjunction with other data collection sources to ascertain psychological profiles of cyberattackers. Once an attacker has initiated an interaction with the HackBot, the tool will be used to establish attack patterns and associated psychological markers (Grasmick et al., 1993). Ten such markers are listed below. While this list is not exhaustive, it nevertheless provides good grounds for early discussions and debate of the utility of HackBots.

1. **Trust bias**: the basis for the establishment of social and collaborative activities (Luhmann, 1988). The psychological default of establishing trust during social interactions is a primary means by which cyberattackers and other nefarious actors are able to manipulate and exploit victims (Lyons & Mehta, 1997).

   Hypothetical HackBot deployment: seek to counter social engineering by engaging the cyberattacker in a simulated dialogue while signaling affiliative and non-threatening motives in order to build trust and extract intel from the cyberattacker. (See Table 1 below).

2. **Online disinhibition**: Suler (2004) argues users may be more likely to engage in disinhibited behavior online, identifying contributing factors that may lead to this behavior, for example; invisibility (perceived anonymity/concealment of identity); asynchronicity and minimization of status and authority in cyber contexts.

   Hypothetical HackBot deployment: seek to inculcate perception of de-anonymization in a synchronous context while manifesting authoritative status.

3. **Impulsivity**: "defined as a personality trait characterized by the urge to act spontaneously without reflecting on an action or its consequences, this trait has been attributed to important psychological processes and behaviors, including self-regulation, risk-taking, and poor decision-making" (Schell, 2020, p. 689) and is thought to underlie various clinical conditions and (due to low risk aversion) is associated with cybercriminal behavior (Aiken et al., 2016; Schell, 2020). Impulsivity is thought to be an underlying component of ADHD, Borderline Personality Disorder and Impulse Control Disorders (Coutlee et al., 2014).

   Hypothetical HackBot deployment: seek to trigger and exacerbate impulsive traits in cyberattackers to induce poor decision making, frustration, burnout and associated decline in attack performance.

4. **Risk taking**: any consciously or non-consciously controlled behavior with a perceived uncertainty about its outcome, and/or about its possible benefits or costs for the physical, economic or psycho-social well-being of oneself or others (Trimpop, 1994). Those who struggle with anxiety and depression are more likely to engage in excessive risk-taking, additionally significant correlations have been established between online delinquency and risk-taking (Brewer et al., 2018).

   Hypothetical HackBot deployment: seek to trigger or exacerbate risk-taking tendencies in cyberattackers, inducing anxiety, errors, frustration and potentially facilitating attribution.

5. **Cognitive overload**: a phenomenon in which someone is given too much information to process at once, or too many simultaneous tasks, thereby hindering performance or processing of the information (Schimming, 2022). Sources of cognitive overload can associated with the complexity of the information or topic itself (i.e., intrinsic); or refer to the manner and format in which information is presented (i.e., extraneous), or to the effort applied to creating a mental representation or schema of the information (i.e., germane) (Sweller, 2011)

   Hypothetical HackBot deployment present misinformation in a format or manner that extracts expenditures of cognitive effort and resources from the cyberattacker, thereby inducing cognitive fatigue, errors and stress.

6. **Reward seeking**: Drawing on the operant and classical conditioning literature, Panksepp observed that the "seeking" cognitive system, mediated by the dopaminergic and mesolimbic features of the central system, is the fundamental motivational force that drives humans to seek information, engage in inquisitive actions, investigate one's environment, and form expectations about future events (Panksepp, 2004; see also Aiken, 2016, p. 52).

   Hypothetical HackBot deployment: exploit the reward-seeking system of the attacker via an intermittent reinforcement learning schedule to reinforce goal-directed cognition and behaviors during dialogue interchange to induce non-productive, compulsive, and perseverant behaviors.

7. **Paraphilias**: a diagnosis of voyeurism disorder requires the offender to be an adult, however voyeuristic tendencies and behaviors are gendered (male dominant) with 50% engaging in voyeurism before the age of 15 (Kaylor & Jeglic, 2021). Qualitative studies have delivered insights from hackers who describe the "seduction of cybercrime," themes extrapolated as follows: thrill, excitement, addiction, curiosity, and voyeurism (Goldsmith & Wall, 2022, p. 107). Criminal surveillance manifests in the cyber context through the use of Remote Access Trojans (RATs), spyware, stalkerware and/or creepware, voyeuristic urges can cause clinically significant distress and impairment in occupational and other areas of functioning. A recent investigation has uncovered the potential relationship between cyberattacker exploit of choice, for example, RATs and paraphilic type disorders such as voyeurism (Aiken et al., 2023b).

   Hypothetical HackBot deployment: seek to engage and exacerbate discomfort, confound and deter cyberattackers, leverage paranoiac tendencies regarding discovery by means of deploying apparent traceability. Advances in HackBot specific NLP technology could simulate an interaction between a victim and a cyberattacker, RATs are typically deployed via phishing attacks and use keystroke monitoring - to counter, the HackBot text may allude to tracking of perpetrator (See Table 1 below).

8. **Dark personality traits**: the Dark Tetrad is composed of four components: narcissism, Machiavellianism, psychopathy, and sadism (Međedović & Petrović, 2015). Dark, anti-social and malevolent traits have been found to be significant predictors of delinquency and criminality (Wright et al., 2017), specifically phishing (Curtis et al., 2018) and hacking (Gaia et al., 2020). Emerging research is exploring the relationship between cyberattacker exploit of choice, for example, Ransomware and Dark personality traits (Aiken et al., 2023b; Martineau et al., 2023).

   Hypothetical HackBot deployment: seek to engage and target impulsivity and need for instant gratification consistent with psychopathic personality type (and hypothetically associated with specific forms of cyberattackers such as those who deploy Ransomware) (see Table 1).

9. **Affective and emotional attributes**: Compulsive and obsessive behavior, particularly among young men (e.g., compulsive online gaming), is often attributed to affective disorders such as depression and anxiety (Aiken, 2016, p. 52).

   Hypothetical HackBot deployment: seek to heighten anxiety and frustration in order to neutralize effectiveness and functionality of cyberattacks.

10. **Attentional tunneling**: refers to "the allocation of attention to a particular channel of information, diagnostic hypothesis or task goal, for a duration that is longer than optimal, given the expected cost of neglecting events on other channels, failing to consider other hypotheses, or failing to perform other tasks" (Régis et al., 2014, p. 1).

    Hypothetical HackBot deployment: present highly salient information in order to siphon and disrupt attentional and cognitive resources.

By way of example, Table 1 below lists cyberattacks, hypothetical cyberattacker vulnerabilities, and (subject to testing) programmed HackBot responses.

Cyberattack	**Cyberattacker Vulnerability**	**HackBot Response**
Phishing	Need to build trust	Engage & extract intel in trust building exchange - use to counter attack
Malware-Spyware	Paraphilic type state: Voyeurism	Engage & leverage paranoia - imply traceability and attribution
Malware-Ransomware	Dark Tetrad traits	Engage & target impulsivity and need for instant gratification

Table 1. Cyberattack and HackBot response

4. Conclusion

Cyber threats and attacks are increasing in complexity, sophistication and velocity. Therefore, successful cyber defense can no longer be sustained with passive defensive tactics. Innovative cyber defense options must be explored in order to execute active defense. This position paper argues for a paradigm shift from passive to active forms of defense against cyberattacks by effectively 'hacking back.' Arguably, this paradigm shift is consistent with state-of-the-art thinking in cyberpsychological applications and cyber-defensive technologies. Automation and scale of response are key requirements in terms of active cyber defense, underscoring the need for, and the utility of, large language model (LLM) defensive solutions. This innovation manifests as a "HackBot" - a hypothetical automated strike back technology serving as an effective honeypot for cyberattackers, engaging them in prolonged, deceptive interactions distracting and draining resources, and specifically conceptualized to reverse socially engineered attacks in cyber defense contexts.

The authors did not receive funding from any organizations during the preparation of this manuscript.

References

Ahmad, A., Hadgkiss, J., & Ruighaver, A. (2012). Incident response teams – challenges in supporting the organisational security function. *Computers Security*, *31*(5), 643–652. https://doi.org/https://doi.org/10.1016/j.cose.2012.04.001

Aiken, M. P. (2016). *The cyber effect: A pioneering cyberpsychologist explains how human behaviour changes online*. John Murray Press.

Aiken, M. P., Davidson, J., Kirichenko, A., & Markatos, E. (2023b). Human drivers of cybercrime: A forensic cyberpsychology approach to behavioural profiling [Manuscript in preparation].

Aiken, M. P., Davidson, J., Walrave, M., Ponnett, K., Phillips, K., & Farr, R. (2023a). Intention to hack? applying the theory of planned behavior to youth criminal hacking [Manuscript in preparation].

Aiken, M. P., Davidson, J., & Amann, P. (2016). *Youth pathways into cybercrime*. https://www.mdx.ac.uk/__data/assets/pdf_file/0025/245554/Pathways-White-Paper.pdf (accessed: 06.12.2023)

Aiken, M. P., Farr, R., & Witschi, D. (2022). Cyberchondria, coronavirus and cybercrime: A perfect storm. In *Cyberchondria, health literacy, and the role of media in society's perception of medical information* (pp. 16–34). IGI Global. https://www.igi-global.com/chapter/cyberchondria-coronavirus-and-cybercrime/293431

Aiken, M. P., & Mc Mahon, C. (2014). The cyberpsychology of internet facilitated organised crime. *Europol Organised Crime Threat Assessment Report (iOCTA)*. https://www.europol.europa.eu/content/internet-organised-crime-threat-assesment-iocta. (accessed: 06.13.2023)

Authorities Concerning Military Cyber Operation, 10 U.S.C. §394 (2018).

Beckers, K., Schosser, D., Pape, S., & Schaab, P. (2017). A structured comparison of social engineering intelligence gathering tools. In J. Lopez, S. Fischer-Hübner, & C. Lambrinoudakis (Eds.), *Trust, privacy and security in digital business* (pp. 232–246). Springer International Publishing.

Bergler, M., Tolvanen, J.-P., & Tavakoli, K. R. (2021). Proceedings of the 31th european safety and reliability conference. presented at the esrel2021.

Berinato, S. (2018). Active defense and "hacking back": A primer. *Harvard Business Review*. https://hbr.org/2018/05/active-defense-and-hacking-back-a-primer

Brewer, R., Cale, J., Goldsmith, A., & Holt, T. (2018). Young People, the Internet, and Emerging Pathways into Criminality: A Study of Australian Adolescents. *International Journal of Cyber Criminology*, *12*(1), 115–132. https://doi.org/10.5281/zenodo.1467853

Broeders, D. (2021). Private active cyber defense and (international) cyber security—pushing the line? [tyab010]. *Journal of Cybersecurity*, *7*(1). https://doi.org/10.1093/cybsec/tyab010

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al. (2018). Universal sentence encoder for english. *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations*, 169–174.

Connolly, I., Palmer, M., Barton, H., & Kirwan, G. (2016). *An introduction to cyberpsychology*. Taylor & Francis. https://doi.org/10.4324/9781315741895

Coutlee, C. G., Politzer, C. S., Hoyle, R. H., & Huettel, S. A. (2014). An abbreviated impulsiveness scale constructed through confirmatory factor analysis of the barratt impulsiveness scale version 11. *Archives of Scientific Psychology*, *2*, 1–12. https://doi.org/10.1037/arc0000005

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

Cyber and Infrastructure Security Agency. (2022). *Cisa expands the joint cyber defense collaborative to include industrial control systems industry expertise*. https://www.cisa.gov/news-events/news/cisa-expands-joint-cyber-defense-collaborative-include-industrial-control-systems (accessed: 06.12.2023)

Deloitte. (2023). *2023 global future of cyber survey*. https://www.deloitte.com/global/en/services/risk-advisory/content/future-of-cyber.html (accessed: 06.12.2023)

Djekic, M. D. (2023). Cyber attack as an asymmetric threat. *Cyber Defense Magazine*. https://www.cyberdefensemagazine.com/cyber-attack-as-an-asymmetric-threat/. (accessed: 06.13.2023)

Fendorf, K., & White, N. (2023). *It's time for the united states to adopt a new strategy to combat ransomware*. https://nationalinterest.org/blog/techland/it's-time-united-states-adopt-new-strategy-combat-ransomware-206508 (accessed: 06.13.2023)

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 (USENIX Security 21)*, 1127–1144. https://www.usenix.org/conference/usenixsecurity21/presentation/ferguson-walter

Gaia, J., Ramamurthy, B., Sanders, G., Sanders, S., Upadhyaya, S., Wang, X., & Yoo, C. (2020). Psychological profiling of hacking potential.

Goldsmith, A., & Wall, D. S. (2022). The seductions of cybercrime: Adolescence and the thrills of digital transgression. *European Journal of Criminology*, *19*(1), 98–117.

Grasmick, H. G., Tittle, C. R., Robert J. Bursik, J., & Arneklev, B. J. (1993). Testing the core empirical implications of gottfredson and hirschi's general theory of crime. *Journal of Research in Crime and Delinquency*, *30*(1), 5–29. https://doi.org/10.1177/0022427893030001002

Groll, E. (2023). *Us intelligence research agency examines cyber psychology to outwit criminal hackers*. https://cyberscoop.com/iarpa-cyber-psychology-hackers/ (accessed: 06.12.2023)

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for nlp. *International Conference on Machine Learning*, 2790–2799.

IARPA. (2023). *Reimagining security with cyberpsychology-informed network defenses (rescind)*. https://www.iarpa.gov/research-programs/rescind (accessed: 06.13.2023)

IBM Security. (2022). *Cost of a data breach report 2022*. https://www.ibm.com/downloads/cas/3R8N1DZJ (accessed: 06.13.2023)

Janosek, D. (2021). What is Cyber Leadership? The Case Study of the 2021 Hacking of a Florida Water Treatment Plant. *United States Cybersecuirty Magazine*, *10*(30).

Kaylor, L., & Jeglic, E. L. (2021). Non-contact paraphilic disorders and offending. *Sexual deviance: Understanding and managing deviant sexual interests and paraphilic disorders*, 171.

Luhmann, N. (1988). Law as a social system. *Northwestern University Law Review*, *83*(1-2), 136–150.

Lyons, B., & Mehta, J. (1997). Contracts, opportunism and trust: self-interest and social orientation. *Cambridge Journal of Economics*, *21*(2), 239–257. https://doi.org/10.1093/oxfordjournals.cje.a013668

Martineau, M., Spiridon, E., & Aiken, M. (2023). A comprehensive framework for cyber behavioral analysis based on a systematic review of cyber profiling literature. *Forensic Sciences*, *3*(3), 452–477.

McAlaney, J., Thackray, H., & Taylor, J. (2016). *The social psychology of cybersecurity*. https://www.bps.org.uk/psychologist/social-psychology-cybersecurity (accessed: 06.12.2023)

McLaughlin, K. L. (2011). Cyber attack! is a counter attack warranted? *Inf. Sec. J.: A Global Perspective*, *20*(1), 58–64. https://doi.org/10.1080/19393555.2010.544705

Međedović, J., & Petrović, B. (2015). The Dark Tetrad. *Journal of Individual Differences*.

Montañez, R., Golob, E., & Xu, S. (2020). Human cognition through the lens of social engineering cyberattacks. *Frontiers in psychology*, *11*(1755). https://doi.org/10.3389/fpsyg.2020.01755

Mouton, F., Leenen, L., & Venter, H. (2016). Social engineering attack examples, templates and scenarios. *Computers Security*, *59*, 186–209. https://doi.org/https://doi.org/10.1016/j.cose.2016.03.004

National Cyber Force. (2023). *The national cyber force: Responsible cyber power in practice*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1148278/Responsible_Cyber_Power_in_Practice.pdf (accessed: 06.13.2023)

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Panksepp, J. (2004). *Affective neuroscience: The foundations of human and animal emotions*. Oxford University Press.

Phillips, K., Davidson, J. C., Farr, R. R., Burkhardt, C., Caneppele, S., & Aiken, M. P. (2022). Conceptualizing cybercrime: Definitions, typologies and taxonomies. *Forensic Sciences*, *2*(2), 379–398. https://doi.org/10.3390/forensicsci2020028

Proofpoint. (2023). *What is social engineering?* https://www.proofpoint.com/us/threat-reference/social-engineering (accessed: 06.13.2023)

Régis, N., Dehais, F., Rachelson, E., Thooris, C., Pizziol, S., Causse, M., & Tessier, C. (2014). Formal detection of attentional tunneling in human operator–automation interactions. *IEEE Transactions on Human-Machine Systems*, *44*(3), 326–336.

Schell, B. (2020). Internet addiction and cybercrime. In *The palgrave handbook of international cybercrime and cyberdeviance* (pp. 679–703). Palgrave Macmillan. https://doi.org/10.1007/978-3-319-78440-3

Schimming, C. (2022). Cognitive overload: When processing information becomes a problem. *Mayo Clinic Health System*. https://www.mayoclinichealthsystem.org/hometown-health/speaking-of-health/cognitive-overload. (accessed: 06.14.2023)

Schmitt, M. N. (2017). *Tallinn manual 2.0 on the international law applicable to cyber operations* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781316822524

Sciacovelli, A. L. (2022). Offensive cyber defense, what are the legal aspects? *CybersecurityItalia*. https://www.cybersecitalia.it/offensive-cyber-defense-which-are-the-legal-aspects/19942/. (accessed: 06.14.2023)

Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., Sap, M., & Shwartz, V. (2023). Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*.

Spitaletta, J. A. (2021). Operational cyberpsychology: Adapting a special operations model for cyber operations (N. I. Ali Jafri, Ed.).

Staal, M., & Stephenson, J. (2013). Operational psychology post-9/11: A decade of evolution. *Military Psychology*, *25*, 93. https://doi.org/10.1037/h0094951

Suler, J. (2004). The online disinhibition effect. *CyberPyschology & Behavior*, *7*(3), 321–326. https://doi.org/10.1089/1094931041291295

Sweller, J. (2011). Chapter two - cognitive load theory. In J. P. Mestre & B. H. Ross (Eds.). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-387691-1.00002-8

Tenney, I., Das, D., & Pavlick, E. (2019). Bert rediscovers the classical nlp pipeline. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4593–4601.

Trimpop, R. M. (Ed.). (1994). Chapter 1: What is risk taking behavior'. In *The psychology of risk taking behavior* (pp. 1–14). North-Holland. https://doi.org/https://doi.org/10.1016/S0166-4115(08)61295-9

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Willett, M. (2023). *Offensive cyber and the responsible use of cyber power*. https://www.iiss.org/en/online-analysis/online-analysis/2023/03/offensive-cyber-and-the-responsible-use-of-cyber-power/ (accessed: 06.13.2023)

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 conference on empirical* *methods in natural language processing: system demonstrations*, 38–45.

Wright, J. P., Morgan, M. A., Almeida, P. R., Almosaed, N. F., Moghrabi, S. S., & Bashatah, F. S. (2017). Malevolent forces: Self-control, the dark triad, and crime. *Youth Violence and Juvenile Justice*, *15*(2), 191–215.

Zaccaro, S. J., Dalal, R. S., Tetrick, L. E., & Steinke, J. (2016). Psychosocial dynamics of cyber security. https://doi.org/10.4324/9781315796352