Seminars in Proactive Artificial Intelligence for Cybersecurity (SPAIC): Consulting and Research

Ehsan Sheyabni*[1] and Giti Javidi[2]

[1]Information Systems and Decision Science, University of South Florida
[2]Information Technology, University of South Florida

[1]sheybani@usf.edu, [2]javidi@usf.edu

Abstract

The authors have designed a platform for research and consulting through a high-level collaborative seminar series to promote networking in proactive artificial intelligence (AI) for cybersecurity (SPAIC). The primary objective is to cover a wide range of techniques in cyber threat intelligence gathering from various social media to dark-net and deep-net, hacker forum discussions, and malicious hacking. The secondary objective is to bring together researchers and consultants in the field to come up with automated and advanced methods of attack vector recognition and isolation using AI and machine learning (ML). In most cases, the hidden nature of security issues makes it hard for fixes in real time. Advanced AI techniques have proven to be superior to the current static methods in cyber threat detection. There have been numerous recent advances in the field of AI, especially in algorithmic approaches such as Speech and Signal Processing, Machine and Deep Learning, Computer Vision, Robotics, Data Mining, Augmented/Virtual Reality, Blockchain, and Cognitive Computing. These highly advanced methods provide tremendous opportunities for behavior/trend based automated analysis, detection, and prevention of cyber attacks/threats. In addition to the potential of development of concepts and whitepapers for a large-scale center, the seminar series will result in identification and recruitment of industrial, academic and/or government partnerships in support of initiatives and research and consulting collaborations as well as creation and support of resources such as research consortia, collaboration sites or social networking tools to facilitate large-scale inter-university research programs in AI and ML in cybersecurity.

Keywords: *Artificial Intelligence, Machine Learning, Cybersecurity, Research Design, and Consulting.*

1. Introduction

The authors have designed and developed a high-level collaborative seminar series to promote networking in proactive artificial intelligence (AI) for cybersecurity (SPAIC). This is a unique and practical model based on many years of authors' experiences and expertise in research, teaching, and professional

*Corresponding Author

service in the field of artificial intelligence for cybersecurity. The intent is not to offer scientific significance of the model, rather present a practical working prototype for research and consulting. In comparison to other similar conferences and/or seminar series ("Artificial Intelligence and Cybersecurity: Attacking and Defending - ISACA Now", 2019), the proposed seminar series are more on the practical side of the spectrum. While other meetings offer pure artificial intelligence with applications in cybersecurity or pure cybersecurity with artificial intelligence as an added flavor, this seminar series delves into the ways cybersecurity and AI are intertwined together and takes a deeper look at the way that solutions and preventive techniques can be generated. In other words, this seminar series is a true motivation for research and consulting as opposed to other seminars where only presentation is emphasized. The primary goals of the SPAIC include:

- Present a wide range of techniques in cyber threat intelligence gathering from various social media to dark-net and deep-net, hacker forum discussions, and malicious hacking.
- Bring together experts in the field to come up with automated and advanced methods of attack vector recognition and isolation using AI and machine learning (ML) ("Artificial Intelligence and Cybersecurity: Attacking and Defending - ISACA Now", 2019).
- Augment the seminar series with various data mining and machine learning techniques as they have proven to recall malicious hacking with high precision.
- Develop concepts and whitepapers for a large-scale center:

    - Identify and recruit industrial, academic and/or government partnerships in support of research and consulting initiatives and collaborations.
    - Create and support resources such as research consortia, collaboration sites or social networking tools to facilitate large-scale inter-university research and consulting programs in AI and ML in cybersecurity.

2. Background

The vulnerabilities of software and the Internet and the accumulation of unprocessed information in big data with many security issues hidden from the cyber security community for many years are serious problems in cybersecurity. There have been many well-documented intrusions into this environment in the recent years, and many zero-day defects are yet to be exposed ("Artificial Intelligence and Cybersecurity: Attacking and Defending - ISACA Now", 2019), (M.A. & C.D., 2018), (Zhou, Zomaya, Li & Ruchkin, 2018). Hence there is a major need for an advanced and comprehensive approach to threat detection, prediction, prevention, and analysis. In recent years, there has been a major research and consulting explosion in the deep learning area, in terms of the number and quality of tools and methods available for predictive analytics (Shickel, Tighe, Bihorac & Rashidi, 2018). Additionally, large organizations are moving their systems onto virtualized platforms or cloud, due to the huge cost savings in hardware, software, maintenance dollars, and to compensate for the lack of skilled workforce. Therefore, a research and consulting platform that introduces these analytical algorithms and systems in virtualized environments would be beneficial in detecting, predicting, preventing, and analyzing cyber threats. The proposed seminar series, SPAIC, bring together experts in the fields of AI in cybersecurity to create an in-depth conversation about robust solutions stemming from research and consulting.

3. Proposed Framework

The authors have performed research in consultation with collaborators to develop, and deploy the SPAIC project. The project consists of 9 seminars in artificial intelligence research and consulting topics related to cybersecurity. The seminars are designed to present the topics from a high-level but collaborative perspective. While specific about intelligence gathering methodologies, the seminars are prescriptive in terms of preserving integrity of systems under cyber-attack and solutions that lend themselves to preventive measures. Each seminar in the series offers an internationally renowned academic or practitioner in the field that not only covers the topic in enough detail, but also makes the connection between the specific topic and proactive techniques for cybersecurity. Each seminar is then followed by several research and consulting activities to 1) engage the audience in a deeper understanding of the topic, 2) create special interest groups (SIG) and community of practice (COP) in that topic, 3) establish a creative environment that would lead into creating effective solutions.

The SPAIC seminar series are designed to cover 9 AI-related topics in cybersecurity. While these seminars are broadcasted (as webinar) and published as open access proceedings, journal articles, and book chapters, it is anticipated that some practical solutions will evolve and get disseminated as a result of discussions, research, consulting, and collaborations in these seminars. What follows is a brief description of the SPAIC seminar series and their relevance to research and consulting in cybersecurity:

3.1. AI Basics

The synergy between AI, new technologies for cybersecurity, and new physical hardware is essential to better understanding of urgent challenges central to the Internet of Things (IoT) and Smart Cities (SC). This smart foundation would be in charge of controlling critical infrastructure, such as CCTV security networks, electric grids, water networks, and transportation systems. Without the continuous, reliable functioning of these assets, economic and social disruption will ensue. To deploy a model for organizing the IoT and SC, AI can be applied to the electric power grid, so as to get maximum benefit from the new technologies. This new grid, called the “4th generation intelligent grid” would use intelligent system-wide optimization to allow up to 80% of electricity to come from renewable sources and 80% of cars to be pluggable electric vehicles (PEV) without compromising reliability, and at a minimum cost to the Nation. Meanwhile, this provides the highlights of the progress made, the open challenges, and important connections to the larger needs of humanity in this field. Unfortunately, this grid is hackable and difficult to secure from cyber-attacks. This leaves IoT and SC in a state of perpetual uncertainty and the risk that the stability of our lives will be upended. Public administrators do not have a good way of knowing which assets and which components of those assets are at the greatest risk. This is further complicated by the highly technical nature of the tools and techniques required to assess these risks. Using artificial intelligence planning techniques, an automated tool can be developed to evaluate the cyber risks to critical infrastructure. It can be used to automatically identify the adversarial strategies (attack trees) that can compromise these systems. This tool can enable both security novices and specialists to identify attack pathways (Werbos, 2018), (Falco, Viswanathan, Caldera & Shrobe, 2018).

3.2. Speech and Signal Processing

Future network systems will be composed of pervasive and heterogeneous distributions of thousands of hardware and software components that are managed in an unmanned and non-centralized way (Soria Zurita, Colby, Tumer, Hoyle & Tumer, 2017). In these new, complex, and highly interdependent systems, traditional security policies and defense strategies are not effective, as thousands of heterogeneous cyber and physical elements are mixed and connected. New security solutions try to learn about the expected behavior from the system and its components, so if a strange event occurs; adequate preventive, corrective, and/or reactive security actions to detect and stop the potential cyber-physical attack being performed are triggered in an intelligent way. In order to learn about the system and select and apply the adequate security actions, very large datasets containing records of previous behaviors should be analyzed, sometimes in a very fast way. This fact enormously complicates the implementation of these new security solutions, as it is necessary a huge storage capacity, which many domestic systems do not have, and it is needed to work with huge data sets whose processing time prevents making decisions with the required speed. This new paradigm requires applications that can support all these technologies based on pervasive sensing platforms to infer relevant information that is implicit in the acquired data (Bordel, Alcarria, Robles & Martín, 2017). Advanced speech and signal processing techniques can reduce large datasets from sensors and other processing devices, with the objective of enabling new security techniques to detect cyber-attacks in a fast and efficient way based on the calculation of small sets of samples, whose statistical configuration is very similar to the original large dataset. Stochastic models and information theory techniques and theorems can be composed and combined in order to define a mathematical framework.

3.3. Machine and Deep Learning

Over the last few years machine and deep learning (MDL) have migrated from the laboratory to the forefront of operational systems. Amazon, Google and Facebook use machine and deep learning every day to improve customer experiences, suggested purchases or connect people socially with new applications and facilitate personal connections. MDL techniques analyze the behavior of complex data and create an effective model for prediction. MDL depends on multiple layers of artificial neurons forming a large network, which act as the core computing part. During the training phase, MDL uses as many examples as possible to determine the relationship between inputs and the output. The output of the network is compared to the desired output and a gradient descent method is applied to minimize the difference between the actual and computed results. Deep Learning provides automatic feature extraction based on the available data presented and leads to higher accuracy in predicting the output. MDL's powerful capability is also there for cybersecurity. Cybersecurity is positioned to leverage MDL to improve malware detection, triage events, recognize breaches and alert organizations to security issues. MDL can be used to identify advanced targeting and threats such as organization profiling, infrastructure vulnerabilities and potential interdependent vulnerabilities and exploits. MDL can significantly change the cybersecurity landscape. Malware by itself can represent as many as 3 million new samples an hour. Traditional malware detection and malware analysis is unable to pace with new attacks and variants. New attacks and sophisticated malware have been able to bypass network and end-point detection to deliver cyber-attacks at alarming rates. New techniques like MDL must be leveraged to address the growing malware problem (Fraley, 2019), (Gulcehre, 2019).

3.4. Robotics

Recent advances in the field of AI have brought an unprecedented maturity into robotics such that intelligent robots are being developed in the form of autonomous vehicles. The widespread use of these autonomous robots and their potential to be hacked into and do harm has raised serious questions about their security. An example of such security issues is the effective manipulation through an indirect attack on a robotic vehicle using the Q learning algorithm for real-time routing control (Clark, 2019). While there are many factors that are considered in design, development, and manufacture of a smart robot, cybersecurity is not as highly prioritized as it should be. As with other embedded systems a higher priority is placed on development costs and delivering functionality to consumers. As the use of robots continues to grow in the manufacturing, military, medical, eldercare and the automated vehicle markets, greater attention should be given to cybersecurity. There are many current and potential cyber threats to robotics at the hardware, firmware/OS, and application levels with drastic economic and human safety impacts (Kazan, 2016).

3.5. Augmented/Virtual Reality

The developments in AI as well as fast memory devices and microprocessors have also resulted in new and better platforms for virtual and augmented reality (AR/VR). Employing these platforms users find themselves moving through a blend of material spaces and immaterial networks. This invisible layer created from the millions of the data streams and network connections that take place around us tend to get denser with the recent development and deployment of the IoT devices in the urban space. The available technology of Mixed Reality spectrum can be applied to provide an immerse view of the information that exist within the invisible layer of the “cyberspace”. “VR Binoculars”, a digital visualization framework that operates in real time, can be used as a medium to unveil the information that exist in our surrounding space through VR/AR. Specifically, the user is situated within an environment where the digital data visualizations and the physical space are matched together, providing to the user the ability to interact, orient themselves and navigate naturally with the cyberspace environment. This framework promotes a better understanding of the IoT ecosystem, justifies the use of sensors in the public space, and raises awareness about privacy and data sharing (Greunke & Sadagic, 2016).

3.6. Natural Language Processing

Much of the data within the realms of cybersecurity is textual in nature. Traditional internal network devices (e.g., Intrusion Detection Systems (IDS), Intrusion Prevention Systems (IPS), databases, workstations, routers, etc.), hacker community data sources (e.g., hacker forums, Internet-Relay-Chat, carding shops, and DarkNet Marketplaces), Open Source Intelligence (OSINT) sources (e.g., Facebook, Twitter, PasteBin, Shodan, etc.) are ripe with rich information that can significantly aide organizations in developing comprehensive and holistic cyber defenses. Indeed, many cybersecurity companies such as FireEye, Splunk, IBM, Webroot, and many others are looking beyond traditional structured data to mine novel insights out of these rich textual data sources. A common paradigm which many companies and researchers adopt is natural language processing (NLP). To date, numerous traditional NLP and text mining methodologies have been employed for malware analysis, phishing email detection, anomaly detection, and other cybersecurity analytics. These include semantic matching, co-reference resolution, named entity recognition (NER), entity resolution, feature selection, feature reduction, ontology development, topic modelling (e.g., latent dirichlet allocation, latent semantic analysis), and others. In recent years, there has been a shift to emerging deep learning based NLP methodologies. These include neural Information Retrieval (neural IR), language modelling, diachronic linguistics, deep structured semantic modelling, and others. Despite remarkable advances, the unique characteristics of cybersecurity data necessitates the development of novel NLP and text processing methodologies.

3.7.Data Mining

Data mining methodologies such as regression, classification, clustering, and association rule mining has traditionally been used on transactional data generated from businesses. While in its nascent stages within the cybersecurity domain, data mining holds significant promise in advancing numerous traditional analytics. These include malware analysis, IP reputation services, phishing email detection, event correlation, anomaly detection, and others. Data mining can assist in two aspects. First it can help organizations and researchers identify patterns within datasets which are not readily apparent by other analytics approaches (e.g., summary statistics, manual inspection, etc.). Second, it can assist in sifting through large amounts of data in an efficient manner. In a context where the amount of data being generated at staggering rates, these benefits are critical to ensuring that an organization is able to effectively extract key insights from all collected data. Beyond enhancing the aforementioned traditional CTI analytics, data mining can provide an array of new inquiries for cybersecurity. These include, but are not limited to, clustering similar types of network events together, grouping similar threat actors in hacker community platforms (e.g., hacker forums), categorizing log files, detecting an adversary’s tactics, techniques, and procedures (TTPs), stream analytics for live cyber threat intelligence data feeds, and many others.

3.8. Blockchain

With the accelerated iteration of technological innovation, blockchain has rapidly become one of the hottest Internet technologies in recent years. As a decentralized and distributed data management solution, blockchain has restored the definition of trust by the embedded cryptography and consensus mechanism, thus providing security, anonymity and data integrity without the need of any third party. The blockchain technology has played a major role in strengthening security in the Internet of Things (IoT). From a security standpoint, blockchain-based solutions could be, in many aspects, superior to the current IoT ecosystem, which relies mainly on centralized cloud servers. Blockchain's decentralized nature results in a low susceptibility to manipulation and forgery by malicious participants. Blockchain-based identity and access management systems can address some of the key challenges associated with IoT security. Blockchain also plays an important role in tracking the sources of insecurity in supply chains related to IoT devices. Using blockchain, it is also possible to contain an IoT security breach in a targeted way after it is discovered. However, there still exists some technical challenges and limitations in blockchain application in cybersecurity. Adopting attribute-based encryption methods seem to enhance access control strategy (Kshetri, 2017), (Gountia, 2019).

3.9. Cognitive Computing

Advanced AI algorithms are responsible for data triage to identify the true "signals" from a large volume of noisy alerts and "connect the dots" to answer certain higher-level questions about the attack activities at the Security Operation Centers (SOCs). Data triage automatons are normally generated directly from cybersecurity analysts' operation traces. Existing methods for generating data triage automatons, including Security Information and Event Management systems (SIEMs), require event correlation rules to be generated by dedicated manual effort from expert analysts. To save analysts' workloads, data triage rules out of cybersecurity analysts' operation traces are mined and used to construct data triage automatons. This approach makes reduces the cost of data triage automaton generation. A study of the cases shows that this system can use the analysts' operation traces as input and automatically generate a corresponding state machine for data triage. Meanwhile, false positive and false negative rates can be calculated to evaluate the performance of the data triage state machine by comparing with the ground truth (Zhuhadar & Ciampa, 2019).

4. Results and Outcomes

The secondary objective of this seminar series is to bring together experts in the research and consulting fields to come up with automated and advanced methods of attack vector recognition and isolation using AI and machine learning (ML). Obviously, this is not achievable only by providing the audience with relevant seminars. There needs to be more interaction and activity to bring together the experts and encourage them to collaborate on specific topics of interest.

A series of cutting-edge seminars in AI topics for cybersecurity are designed.

These executive cybersecurity briefings are deigned to be at the speed of business. Topics and set up would allow for the involvement of various experts consulting and research and also facilitate the development of a library of seminars that can be maintained for future offerings. The results and outcomes will be shared with universities, industry, and other constituents to help them pursue further research and enhancements.

This research aims to bring together research and consulting experts in AI and ML for cyber threat/attack detection through in-depth seminal series and discussions. Instead of relying on current 'fixed' approaches to cyber breaches, this research relies on a discussion of advanced algorithmic approaches to flexibly detect intrusions. These approaches will open doors to a potential merger of the latest advancements in AI and ML for automated analysis, detection, and prevention of cyber attacks/threats, resulting in advancing the field of cybersecurity. During these seminars, the PIs will engage the participants in discussions and activities with an expectation to make significant advances in security methodologies using AI and ML.

5. Conclusion

Cyber threats, attacks, hacks, and breaches have become a normal incident in day-to-day life of Internet users. The proposed seminar series focus on presenting the cybersecurity community with applied AI and ML algorithms. While these presentations will be broadcasted (as webinar) and published as open access proceedings, journal articles, and book chapters, it is anticipated that some practical solutions will evolve and get disseminated as a result of discussions and collaborations in these seminars. Furthermore, the technologies developed as a result of this seminar series has the potential to grow into a larger proposal/project supported by the Department of Defense (DOD), Department of Homeland Security (DHS), and National Science Foundation (NSF) providing performant and scalable solutions to government and private entities. The seminar series focus on techniques developed to identify emerging cyber threats including information on newly developed malware and exploits that have not yet been deployed in a cyber-attack. The seminar series will be augmented with various data mining and machine learning techniques as they have proven to recall malicious hacking with high precision. The proposed ideas are unique and novel in the area of cybersecurity and with sufficient time and effort can result in the development of an extremely powerful tool for early threat detection and defense. The proposed research and its findings can be shared with teams working in similar areas in other universities and educational institutions to further advance the field of cybersecurity.

References

Artificial Intelligence and Cybersecurity: Attacking and Defending - ISACA Now. (2019). Retrieved from http://www.isaca.org/Knowledge-Center/Blog/Lists/Posts/Post.aspx?ID=1157

Bordel, B., Alcarria, R., Robles, T., & Martín, D. (2017). Cyber–physical systems: Extending pervasive sensing from control theory to the Internet of Things. *Pervasive And Mobile Computing, 40*, 156-184. doi: 10.1016/j.pmcj.2017.06.011

Clark, G. (2019). A Malicious Attack on the Machine Learning Policy of a Robotic System (pp. 516 – 521).

New York, NY, USA: Proceedings of 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications.

Falco, G., Viswanathan, A., Caldera, C., & Shrobe, H. (2018). A Master Attack Methodology for an AI-Based Automated Attack Planner for Smart Cities. *IEEE Access*, *6*, 48360-48373. doi: 10.1109/access.2018.2867556

Fraley, C. (2019). The promise of machine learning in cybersecurity - IEEE Conference Publication. Retrieved from https://ieeexplore.ieee.org/document/7925283/

Gountia, D. (2019). Towards Scalability Trade-off and Security Issues in State-of-the-art Blockchain. *ICST Transactions On Security And Safety*, *5*(18), 157416. doi: 10.4108/eai.8-4-2019.157416

Greunke, L., & Sadagic, A. (2016). Taking Immersive VR Leap in Training of Landing Signal Officers. *IEEE Transactions On Visualization And Computer Graphics*, *22*(4), 1482-1491. doi: 10.1109/tvcg.2016.2518098

Gulcehre, c. (2019). Deep Learning. Retrieved from http://deeplearning.net/

Kazan, H. (2016). Contemporary Issues in Cybersecurity. *Journal Of Cybersecurity Research (JCR)*, *1*(1), 1. doi: 10.19030/jcr.v1i1.9745

Kshetri, N. (2017). Can Blockchain Strengthen the Internet of Things?. *IT Professional*, *19*(4), 68-72. doi: 10.1109/mitp.2017.3051335

M.A., A., & C.D., J. (2018). Automated multi-level malware detection system based on reconstructed semantic view of executables using machine learning techniques at VMM. *Future Generation Computer Systems*, *79*, 431-446. doi: 10.1016/j.future.2017.06.002

Shickel, B., Tighe, P., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE Journal Of Biomedical And Health Informatics*, *22*(5), 1589-1604. doi: 10.1109/jbhi.2017.2767063

Soria Zurita, N., Colby, M., Tumer, I., Hoyle, C., & Tumer, K. (2017). Design of Complex Engineered Systems Using Multi-Agent Coordination. *Journal Of Computing And Information Science In Engineering*, *18*(1), 011003. doi: 10.1115/1.4038158

Werbos, P. (2018). AI Intelligence for the Grid 16 Years Later: Progress, Challenges and Lessons for Other Sectors. *Proceedings of 2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1 – 8, 8-13 July 2018, Rio de Janeiro, Brazil.

Zhou, X., Zomaya, A., Li, W., & Ruchkin, I. (2018). Cybermatics: Advanced Strategy and Technology for Cyber-Enabled Systems and Applications. *Future Generation Computer Systems*, *79*, 350-353. doi: 10.1016/j.future.2017.09.052

Zhuhadar, L., & Ciampa, M. (2019). Leveraging learning innovations in cognitive computing with massive data sets: Using the offshore Panama papers leak to discover patterns. *Computers In Human Behavior*, *92*, 507-518. doi: 10.1016/j.chb.2017.12.013