Digital Ghost: An Exploration of Next Generation of AI Cyber Defense Opportunities and Challenges for Defenders of Critical Energy Infrastructure.

Michael Mylrea, Matt Nielsen, Justin John, Masoud Abbaszadeh

Abstract

Successful adoption of next generation machine learning algorithms and AI to empower cyber defenders of our nation's power grid is critical to the security and well-being of any modern economy. Sophisticated cyber adversaries are increasingly deploying technology such as AI combined with stealthy tactics and techniques to attack critical energy infrastructure. Defenders of these modern infrastructures need to better understand how to combine innovative technology in a way that empowers their teams to respond to a complex, non-linear and rapidly evolving cyber threat. The following paper first (i) explores opens exploring how AI is being combined with advances in physics to develop a next gen solution to defend against sophistic cyber-physical attacks to critical infrastructure. (ii) The next section provides an overview of the technology and explores its applicability to address the needs cyber defenders of critical energy infrastructure of the future. Applicability is explored through opportunities and challenges related to human-machine team or people as well the process and technology (iii) The next section will include validation and verification findings when the technology was tested defending against stealthy attacks on the world's largest gas turbine. (iv) Then the article explores how the AI algorithms are being developed to provide cyber defenders with improved cyber situation awareness and rapidly detect, locate and neutralize the threat. (v) The article concludes with future research and overcoming human-machine challenges with neutralizing threats from all hazards.

Introduction

A digital transformation of critical energy infrastructure is underway, digitizing, networking and automating the energy value chain. Today's smart energy system are increasingly interoperable, two-way, agile and flexible in incorporating distributed energy resources that have helped transition energy usage and consumption to lower carbon, sustainable, renewable energy. However, this digital transformation has also created new cyber-physical security challenges in securing an array of vulnerable energy delivery systems and associated operational technology. The rapid digital transformation of our critical systems has significantly increased its attacks surface by combining cyber-physical systems, software and hardware, information technology (IT) and operational technology (OT). This has created new challenges to identify and protect these critical systems with real-time cyber-physical situational awareness and determine a base's overall cyber threat surface in terms of control systems, automation and other operational technology. Today, most cyber defense and monitoring solutions are reactive, innovative new resilient systems -like Digital Ghost- are more holistic in their defense, ensuring continuous operations via graceful degradation and autonomous recovery.

Cybersecurity Gaps for Advanced Detection, Protection and Monitoring Solutions

Grid modernization has spurred the integration of DER's and electricity infrastructure that is increasingly digitized, networked, automated and complex in its communications using multiple languages and protocols between an increasing number of parties. Securing these critical communications in transit, at rest and at the device level without sacrificing improvement in forecasting, control and optimization of these assets is essential. Indeed, any effective cybersecurity solution should not curtail advances in control and optimization. The figure below highlights how Grid cyber defenders have responded to cyber threats to DERs with various cybersecurity solutions that try to segment and "air gap" critical systems. However, these cybersecurity solutions do not

weather, insider threats, human error, supply chain attacks on software, hardware, etc.

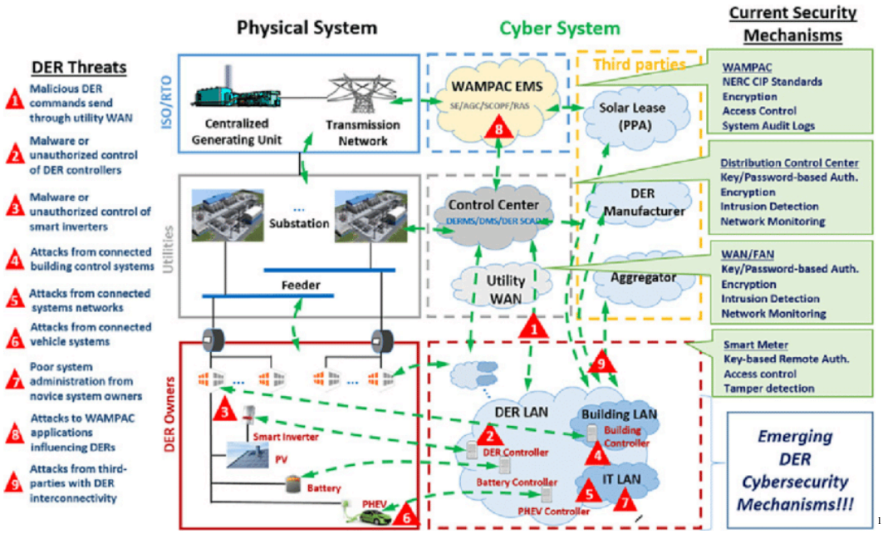

Figure 1. Cyber Physical Threat to Distributed Energy Resources (DERs).

As a result, there are numerous cybersecurity gaps for advanced detection, protection and monitoring of energy delivery systems, networks and interconnected energy delivery systems. These gaps could potentially be exploited to cause degradation of service and potential cascading failures to the power grid. However, due to the many gaps in detection and monitoring it is difficult to quantify the threat and risk. Moreover, increased monitoring and detection of electricity infrastructure may give the perception that attacks to grid are increasing when in fact this is a measure of improved cyber situation awareness. When media suggests there is an increase in cyber-attacks on the grid, is that because monitoring and detection technology approved or because threat groups are increasingly targeting the grid. Currently, there is a major gap in research and data available to quantify these risks. Another major cybersecurity gap for advanced detection, protection and monitoring is found with the increasing penetration of distributed energy resources (DERs).[2,3] Increased connectivity and two-way communications of DERs with infrastructure associated with the bulk power grid will require advanced threat monitoring and detection to address existing and potential future cybersecurity gaps.[4] Any holistic solution requires a comprehensive approach of human and machine or people, process and technology. But many gaps remain.

Policy Gaps - Currently the NERC CIP cybersecurity requirements have increased defenses for critical systems found in the bulk grid. However, distribution and grid edge devices that are increasingly connected to bulk grind infrastructure are vulnerable to sophisticated cyber-attacks.

Technology Gaps - The data and connectivity requirements needed to improve grid edge and DER management - increased awareness, control to level loads and manage capacity constraints and reverse power flow – has significantly expanded the attack surface of our nation's grid. For example, solar energy systems grid-support functions can manipulate to diminish reliability and damage electricity infrastructure. Securing PV systems critical communications at rest as well in transit to

[1] Qi, Junjian & Hahn, Adam & Lu, Xiaonan & Wang, Jianhui & Liu, Chen-Ching. (2016). Cybersecurity for Distributed Energy Resources and Smart Inverters. IET Cyber-Physical Systems: Theory & Applications. 1. 28-39. 10.1049/iet-cps.2016.0018.
[2] https://www.utilitydive.com/news/security-and-distributed-resources-an-attacker-will-eventually-get-in-s/565966/
[3] Greenberg, Andy, Emily Dreyfuss, Brian Barrett, Danny Gold, Issie Lapowsky, and Lily Hay Newman. "How Hacked Water Heaters Could Trigger Mass Blackouts." Wired (Aug. 2018) (2018).
[4] Lee, A. "Electric sector failure scenarios and impact analyses." National Electric Sector Cybersecurity Organization Resource (NESCOR) Technical Working Group 1 (2013).

challenging due to the increased internet connectivity and digitization [5] as well as communication protocols that prioritize interoperability but lack basic encryption and authentication mechanisms.[6]

Together, current policies, processes and technologies that prioritize interoperability and connectivity and do not provide high fidelity cyber situational awareness to detect cyber-physical anomalies to DERs. Even when monitoring is available, determining the cause of the anomaly and localizing and neutralizing the threat is a major gap in this space. Sophisticated adversaries can perturb systems to instigate abnormal power flow, supply chain attacks can push updates to behind the meter systems to add or drop load in a way that could potentially cause a grid level even, insider attack can cause instabilities like sub- synchronous resonance, man in middle attack amplify weak grid conditions, just to name a few.

Digital Ghost: A Next Generation Response to Close Critical Energy Infrastructure Gaps

In response, researchers at GE Global Research, in partnership with the U.S. energy industry and the U.S. Department of Energy, have developed innovative solution to identify, mitigate and autonomously respond to evolving cyber threats. This next generation cyber-physical anomaly detection solution combines advances in machine learning (AI) to rapidly identify, protect, detect, respond and recover to cyber-physical threats and vulnerabilities targeting operational technology (OT). If an adversary attacks, manipulates or compromises a system integral, Digital Ghost helps detect anomalous behavior, locate and neutralize the attack while maintaining availability and integrity of critical operations. To realize this goal, Digital Ghost leverages machine learning of digital twins (high-resolution models of OT/IT systems and networks) in order to: ***Identify***, detect and map critical systems and detect anomalies and associated vulnerabilities and quantify them; ***Localize***, Isolate and Protect critical controls systems and OT (sensors/actuators/drives/controllers) and analyze the anomaly, and; ***Neutralize*** to autonomously ***Respond*** and ***Recover***, mitigating advanced threats. The ability to review the control logic and autonomously maintain operations without losing availability of critical systems is a potential game changer to cyber-physical resilience, but many challenges remain

Cyber defense of critical infrastructure continues to evolve, but cyber adversaries often have the upper hand as offensive tools improve and the attack surface expands. Cyber challenges and gaps to policies, technology and people (workforce and expertise). To change this equation, new paradigms and formal methods as well as advances in threat mitigation technology need to be developed. Even as cyber defense technology approves, workforce development especially in the area of OT cybersecurity remains a major gap. The confidentiality, integrity and availability triad that has defined cybersecurity in the last 20 years continues to be pressured by the digital transformation underway that prioritizes interoperability, connectivity and a move towards automation. As we digitize, automate and connect systems in critical infrastructure to the internet this also expand the cyber-physical attack surface.

To improve the current state of the art in grid cyber defense requires moving beyond the cybersecurity triad paradigm to cyber resilience, which assumes we can identify, detect, respond and recover to cyber threats and vulnerabilities in sub-second times. Cyber resilience includes a hardened perimeter but also neutralizes sophisticated attacks once they have found.
Advances of innovative threat mitigation solutions helps move the industry towards cyber resilience. However, the design and implementation of these advances, such as machine learning algorithms,

---

[5] Johnson, Jay. *Roadmap for photovoltaic cyber security*. Sandia Tech. Report, SAND2017-13262, 2017.
[6] Onunkwo, Ifeoma, B. Wright, P. Cordeiro, Nicholas Jacobs, Christine Lai, Jay Johnson, Trevor Hutchins et al. *Cybersecurity assessments on emulated DER communication networks*. Sandia Technical Report, 2018.

cyber defense technology needs to be complemented by a process function in a way that turns data into intelligence. Through this information fusion human machine teams can increase both their autonomy and effectiveness to evolve their defenses to be cyber resilience in response to sophisticated evolving threats. The following provides an overview of the design and deployment of a next generation AI cyber defense technology to detect, localize and neutralize threats in a more effective, autonomous way. To realize that goal requires **the leveraging the science of interdependence for autonomous human-machine teams in a synergistic way to identify and overcome existing gaps with** people, process and technology explored below.

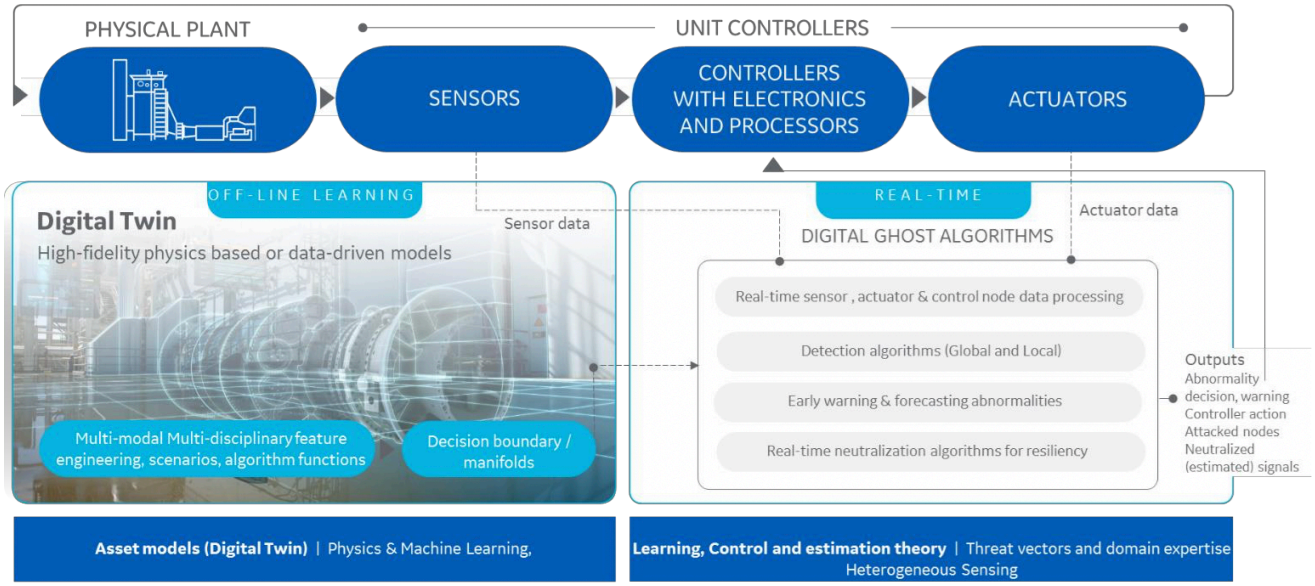

Figure 1: Digital Ghost functionality diagram. Example is of a power generation plant. The top portion in the figures depicts a complex system with sensor, controls and actuators. The bottom left pane shows how the Digital Ghost is trained from off-line operational data and physics based models. The bottom right pane outlines the real-time algorithms providing detection and neutralization functions.

II. People, Process and Technology Applicability & Gap Analysis

This section examines the applicability of existing cybersecurity technology to address cyber defenders needs in modern critical energy infrastructure, which is going through its own digital transformation. Applicability and gap analysis is explored through opportunities and challenges related to human-machine team or *people* as well the *process* and *technology*.

Attack Detection

Attack Detection – Advanced threat detection starts with a comprehensive design. Digital Ghost's design phase started with scoping the target system and defining the sub- systems that are of primary interest. Instead of a purely unsupervised approach to developing the machine learning algorithms, system experts draw on their domain knowledge to establish a matrix of credible cyber-attacks, naturally occurring faults and vulnerabilities in the system. The highest impact abnormalities (ie. Attacks/faults) are chosen for computer model simulation. The high-fidelity Digital Twin models are exercised to define the system operating boundaries. Normal operating space is mapped out as well as attack/abnormal operating space. The machine learning algorithm developed from these defined scenarios is intended to differentiate between a naturally occurring system fault or degradation mode and a likely malicious cyber-attack scenario. Historical data obtained from the asset or plant is reviewed to establish the key system monitoring nodes. An optimization algorithm is then used to establish the decision boundary, called decision manifold, between the normal and attack/fault (abnormal) operating regions. Performance predictions are then generated based on this optimal decision boundary. The optimal decision boundary is also updated over time in the future as the system evolves via real-time learning and adaptation algorithms. The next step is deploying the

detection algorithm performance is reviewed and continuously monitored.

Technology Gaps - (i) Unlike IT solutions which are easy to enumerate and inventory via scanning, operational technology includes a diverse attack surface that is often connected through both internet protocol (IP), serial, and other connections. (ii) Proprietary protocols that are often vulnerable by design as vendors prioritize functionality, ease of use and cost over security. (iii) Firewalls, network and host intrusion detection systems which are limited to defending against malicious signatures that are not in their libraries of attacks signatures. Thus, a brute force, polymorphic, AI generated, or insider attack will be very difficult to detect. Zero-day exploits targeting operational technology are very difficult to block with most existing attack detection solutions that are designed for IT. (iv) Resource intensive tuning can be required for AI defense solutions so it is critical solutions integrate into existing technology stack and Security Information and Event Management (SIEM).

Process and Policy Gaps: As AI solutions improve attack detection it will increase the speed, size and fidelity of logging critical machine state integrity as well as other network and system outputs. Thus, monitoring policies and process updates need to intelligently distill down and fuse these findings for this data to create actionable cyber intelligence. Often, grid cyber defenders have policies and processes in place to monitor and log their critical cyber assets as defined by the NERC CIP requirements; however, they often times don't read these logs. Moreover, additional network or systems that are connected to these critical cyber assets can provide an attack pathway if not secured.

People Gaps: Machine learning algorithms that have high false positive rates create a prohibitive operations and maintenance requirements on security teams. Cybersecurity teams have been traditionally IT focused, however, convergence of IT/OT in critical infrastructures has increased the responsibilities and created new workforce development challenges. Some innovative new tools require training and adding another tool creates information fusion challenges. Finally, AI solutions that are tuned and learn what is normal on networks and systems that are already infected may be providing a false sense of security to their operators. Advances in invariable learning and humble AI explored in this paper highlight how researchers are overcoming these gaps.

Attack Localization

This phase develops a software algorithm to localize the attack to a specific system function. Attack dependency tests are conducted to further separate the attacks into independent or dependent attacks. Local decision manifold boundaries are created for each monitoring node using data sets by running various attack scenarios with the high-fidelity Digital Twin models mentioned previously. The system post-processes the localized attack and determines whether the detected attack is an independent attack or an artifact of the previous attack through propagation of the effects in the closed-loop feedback control system. This provides additional information and insight and is useful when multiple attacks are detected. The same approach is practiced for localization when naturally occurring faults are detected.

Technology Gaps: For critical OT asset and systems the sub second time requirements for effective detection and localization is a major gap for most cyber defense solutions. Moreover, there is a lack of real-time detection and localization solutions to responds to cyber-attacks. Visibility of data, and probable fault or attack is limited across the energy value chain. Advances in supervisory control and data acquisition as well as energy management and distribution management systems have increased fidelity and control of data. Similarly, advances in active scanning and interrogating/communicating with an OT in its native protocol has increased visibility. However, many gaps remain and have created prohibitive localization response times. Thus, speed of response for malware and infiltration mitigation to an attack is a critical gap that needs to maintain reliable,

IT/OT environments that combine cyber and physical, legacy and modern system and various protocols.

Process and Policy Gaps: Current processes focus on localizing faults, safety and reliability issues. Cybersecurity is often an afterthought. Systems engineering approaches in practice are often reduced to adages, such as if it's not broken, don't fix it. Or even the colloquial KISS expression - Keep it stupid simple. As a result, most policies focus on how to localize and respond to sensor or actuator faults, component level faults, system level faults that could cause loss of power or degradation in output, not how to localize a cyber-attack. Thus, there is a real risk that adversary could imbed themselves into a critical system, establish a stealth command and control channel and potentially carry out an attack undetected at a later date.

Human Resource Gaps: Locating a fault in a complex system of system like a power plant is no trivial task. In addition, the resource gaps noted for detection, localization has similar and related issues related to localizing the actual system that faulted; especially during a transient even or when there is a highly variable stochastic load- events that create a lot of noise and challenges human operators' ability to localize the problem. Moreover, sensor or actuator faults, component level faults, system level faults, and cyber-attacks may all produce similar effects in a system (i.e., loss of power or degradation in output).

Attack Neutralization

This phase is aimed at creating technologies to self-protect and remove the effects of attacks on the monitoring nodes so that the system can continue to function even in the presence of attacks. Two approaches are being investigated to provide robust neutralization to all affected monitoring nodes. One approach uses a featured-based multi-node virtual sensor as a resilient estimator and feedback controllers to neutralize the abnormal behavior in the system caused by the cyber-attacks. The feature- based multi-node virtual sensor provides estimates for affected monitoring nodes using information from the localization module. In the second approach, the virtual sensor is constrained by the decision boundary used during attack detection stage. The estimates of true operational signals for the monitoring nodes are computed on a sample-by-sample basis simultaneously using all data from monitoring nodes. True operational signals are provided to the control system on a continuous basis while informing the operator as and when attacks are detected.[7]

Technology Gaps: For critical OT assets and systems there are sub second time requirements for effective communications. Cyber resilience requires the ability to both detect and localize rapidly to effectively an accurately neutralize an attack or anomaly. Sophisticated cyber-attacks, zero day exploits, hybrid cyber-physical, insider threats, to name a few, create challenges in neutralization. Control systems are designed with functionality, ease of use, safety, cost and connectivity in mind, not security. This creates additional challenges related to neutralization. The TRISIS cyber exploit was exemplarily of these design vulnerabilities where a safety instrumented control system was exploited in a sophisticated attack on operational technology.

Process and Policy Gaps: Today, cyber security policies for critical energy infrastructures often prioritizes availability and integrity of critical systems, however, most current solutions only identify threat and vulnerabilities, relying on manual response; (ii) Manual responses create resource and response time challenges that are prohibitive; (iii) Existing tools lack prioritization and create prohibitive resource requirements with false positives and lack of prioritization.

---

[7] John, J., Nielsen, M, Abazadeh, M, Markam, (2020) Advanced Detection and Accommodation Research Findings. GE Global Research (need to update this citation with names, etc..

Anomaly forecast enables early detection of salient and stealthy attack which could otherwise, remain in the asset for days or even months without being caught. It also enables early engagement of the system operator or the automatic accommodation in a cyber incident. Furthermore, the anomaly forecast system can predict future system failures / malfunctions and be used as a tool for predictive health monitoring and prognostics. Once the security of a system is compromised, the adversarial impact will propagate through the system until it gets detected by the attack detection mechanisms. However, by the time that those mechanisms have detected an attack, the damages may have been already done, and the impact may be too large to be accommodated. This invention, provides early warning capability to the attack detection so that a security breech is detected and alarmed at an early stage both for operator response and for attack accommodation.

The outputs of prediction models in different time-scales (aka, future values of the features) are compared to the corresponding decision boundaries for anomaly forecasting. While comparing the feature vectors to the decision boundary, estimated time to cross the decision boundary will provide information for future anomaly. If a future anomaly is detected, an early warning is generated in the operator display with anticipated time to reach anomalous state and a message is sent to the automatic accommodation system for potential early engagement.

Digital Ghost Research Findings and Future Research

Invariant Learning

Measuring both anomalies and invariances in deep networks for a complex system of system like the power grid is not an easy task. For one increased penetration of stochastic intermittent distributed energy resources further complicates essential pattern recognition tasks to flag anomalies and variances. Recent research on shows advances in training deep architectures in a supervised manner to be invariant to multiple confounding properties and input transformations as found in electricity infrastructure.[8] Future research examining how to improve invariant machine learning to improve cyber-attack detection and accommodation (ADA) accuracy of Digital Twin Models to identify and protect against cyber physical attacks on critical energy systems and infrastructures is essential.

Modeling a complex system of systems in electricity infrastructure is challenging due to a number of issues from bias offsets between the actual values of the key nodes being monitored and those simulated to “noise” on the system. What appears as an anomaly could be caused by human error, computational error, a natural occurring weather and ambient events, increases in supply and demand, a cyber-attack or hybrid cyber-physical event. Moreover, adversaries could potentially exploit continuous machine learning bias with next generation machine learning attacks that slowly bias key nodes such that the continuous learning system “learns” this incorrect behavior and treats it as normal. To overcome these challenges, next generation cyber resilient invariant learning algorithms need to be improved to advance physical detection and mitigate risk from sophisticated AI attacks. Moreover, for these innovative technology solutions to be successfully transitioned to the energy sector will require how alerts of cyber events are best displayed to grid cyber defenders, which are already distracted with many tools, screens and the day to day challenge of keeping the grid reliable and balanced.

Research findings on cyber-attack detection and accommodation (ADA) accuracy of Digital Twin Models to identify and protect against cyber physical attacks provides valuable insight into both

---

[8] Goodfellow, I., Lee, H., Le, Q. V., Saxe, A., & Ng, A. Y. (2009). Measuring invariances in deep networks. In *Advances in neural information processing systems* (pp. 646-654).

experiments was the mismatch between the Digital Twin model of the gas turbine and the actual physical system. For example, oscillations were seen during the early set of test cases, later deemed caused by a poorly tuned control loop. However, the team struggled to reproduce the oscillations in the simulator. During the last set of test runs, the ambient temperature was much colder than anticipated. Due to the cold temperatures, the gas turbine used inlet bleed heating, which was a mode of operation not thoroughly explored during the training phase. Bias offsets between the actual values of the key nodes being monitored and those simulated were the main result.[9]

These findings point towards the need to employ continuous learning to modify the algorithms and/or decision manifolds based upon actual field data. Allowing flexibility for the algorithms to be modified or adjusted based upon actual field data could help alleviate the model mismatch. However, continuous learning could also create a potential new cyber-attack method where an attacker slowly bias key nodes such that the continuous learning system "learns" this incorrect behavior and treats it as normal. Advances in invariant learning are needed to mitigate manipulation of the continuous learning algorithms. Research findings suggest that improvements can be made in part by leveraging the physics of known degradation curves. Thus, the invariant learning module would confine the continuous learning system to only learn performance degradation that has been physically realized. Another set of manifolds could be created that are used to confine or check the continuous learning process. If effective, these controls could also help limit adversarial AI attacks that would bias and poison algorithms.[10]

Autonomous Defense: Critical Sensors Identification & Trust

Self-healing complex systems of systems are the holy grail of cybersecurity research and development. Conference organizers highlight the many challenges many that affect "the design, performance, networks operating autonomous human-machine teams.[11]" Research findings from testing Digital Ghost's neutralization algorithms suggest these challenges increase when human teams lack observability and context on a complex transient system such as a gas turbine. More specially, the neutralization algorithm had difficulty providing an optimal estimate for critical sensor within an acceptable error bounds allowable for closed-loop control. Additional context and domain knowledge from the team helped overcome that challenge and found that this sensor had poor observability from the remaining non-attacked sensors. This suggests that advances in autonomous cyber defense must prioritize observability of remaining non-attacked sensors to calculate an estimate that would work in closed-loop control.

Future research on the science of interdependence for autonomous human-machine teams combined with advances in control theory methods may help improve ability of machine learning algorithms to decide which sensors have poor observability before moving to deployment. In complex transient system of system there is a need to improve observability and trustworthiness of critical energy delivery sensors to autonomously protect, detect and recover or neutralize cyber-physical threats. In absence of these capability, near terms opportunities to improve the state of the art in neutralization, include determining what sensors lack observability for neutralization and create an alert to human operators signaling the inability for neutralization to provide a corrective action if one of these nodes were attacked. Applying advanced encryption and authentication mechanisms for this sensor via

---

[9] John, J., Nielsen, M, Abazadeh, M, Markam, (2020) Advanced Detection and Accommodation Research Findings. GE Global Research (need to update this citation with names, etc..

[10] Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., & Li, J. (2018). Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9185-9193).

[11] W.F. Lawless; Ranjeev Mittu, Don Sofge; Thomas Shortell, and Tom McDermott. AI welcomes Systems Engineering: Towards the science of interdependence for autonomous human-machine teams. AAAI Spring Symposium 2020 Abstract.

would help the information security community better understand how we can improve control theory methods and combine human-machine teams in a way where machine learning algorithms can empower cyber defenders better determine the integrity and trustworthiness of critical sensors?

Digital Twin Adaptive Thresholds

Electricity infrastructure is a complex system of system with cyber-physical convergence that creates challenges in observability, trustworthiness of cyber and physical data. Not surprisingly, Digital Twin model mismatch to the actual physical turbine caused several issues for researchers. When IBH was on, the global detection score was noticeably lower as compared to when this mode was not activated. The team manually lowered the detection threshold to maintain the desired performance. The team's role in creating adaptive thresholds was influenced by deep domain knowledge of operations of the gas turbine being tested and multiple threshold levels were chosen based upon certain operational criteria. These thresholds may be pre-determined with specific activation criteria. Or machine learning models could be programmed during training phase to modify the threshold levels once deployed to the field. Future research should explore the importance of context in human-machines teams as it relates to digital twin methodologies and technologies to give impetus to next generation machine learning and cyber resilient solutions.

Humble AI

Humble AI is making valuable advances in marrying man and machine, answering such questions as: How can the algorithms recognize this fact and alert the operator of a potential decrease in accuracy or confidence in the classification results? How can the ML/AI methods recognize they are being asked to extrapolate into previously unseen operating regions? What is the proper response if this does happen? Does Digital Ghost or other advanced AI cyber defense halt operation? Or does it continue but express some notion of a reduced confidence in results? Next generation AI cyber physical anomaly detection and neutralization require continuous improvement of ML/AI methods that are agile, adaptable and evolve with a complex, non-linear and evolving threat. R&D findings from Digital Ghost algorithms that are trained off-line to create the various decision manifolds for both local and global detection need to be able to adapt to the field operating conditions of all hazards – cyber, physical, naturally occurring – as conditions of critical energy delivery systems move away from the training points and extend into regions previously not simulated. If when in the field operating conditions move away from the training points and extend into regions previously not simulated, it is essential that the algorithms recognize this fact and alert the operator of a potential decrease in accuracy or confidence in the classification results.

Explainable AI

Explainable AI is the concept of being able to understand how the ML/AI algorithms are arriving at their solutions, often through advanced human-machine interfaces containing easy to understand visualization techniques. In this report, we present an attempt at a graphical method to help point towards which features and to some extent what nodes are contributing to the abnormal global detection score. However, more work needs to be done in this space in relation to the ADA technology. The ADA technology contains complex algorithms, some of which are constructed using machine learning, AI techniques. While this demonstration at GE's test stand helps validate the technology, future customers may still have a skepticism because of the complexity and non-intuitiveness contained within the highly non-linear algorithms.

Ghost's AI and Machine Learning (ML) capabilities without reducing fidelity and accuracy of detection, localization and neutralization capabilities. It is essential that Digital Ghost's next generation cyber physical anomaly detection and neutralization algorithm contain technical complexity that are not always intuitive to grid cyber defenders. This creates a number of human and cyber-physical integration challenges that could explored with future research on how best integrate humans and machines. Lessons learned from DG research has helped develop c complex algorithms, some of which are constructed using machine learning, AI techniques. While this demonstration at GE's test stand helps validate the technology, future customers may still have a skepticism because of the complexity and non-intuitiveness contained within the highly non-linear algorithms.

Conclusion

Grid modernization has been accompanied by a digital transformation that has increasingly digitized, networked and automated the energy value chain. Today's smart grid is increasingly two way, agile and flexible in incorporating distributed energy resources that have helped transition to a lower carbon economy. Research highlighted in this paper highlighted how this digital transformation must marry man and machine. Similarly, research findings also suggest that human machine teams can be empowered and also blindsided by AI by given a false sense of security. The "smart" grid has increased connectivity and created new cyber-physical security challenges in securing an array of vulnerable energy delivery systems and associated operational technology. As a manufacturer of a large percentage of the world's power systems, GE has been integral to grid modernization and has unique insight as well as responsibility to ensure more holistic cyber resilient policies, processes and technology.

Realizing this goal is imperative as U.S. electricity infrastructure and will require a holistic approach of people, policies and technology. Research findings suggest successful adoption of next generation technology, such as AI algorithms found in Digital Ghost. Findings also suggest that innovation should not happened with humans out of the loop. The form of the technology R&D must compliment the function and independencies of the team in order to empower cyber defenders of our nation's power grid. This is especially true as sophisticated cyber adversary are increasingly deploying technology such as AI combined with stealthy tactics and techniques to attack critical energy infrastructure. Defenders of these modern infrastructures need to better understand how to combine innovative technology in a way that empowers their teams to respond to a complex, non-linear and rapidly evolving cyber threat. Novel technology advances combining domain expertise in physics and next generation AI solutions will only be successful if humans are empowered in the loop, not disintermediated. This is especially true when defending against a diverse, complex, non-linear and rapidly evolving threat of human adversaries which are executing sophisticated cyber-physical attacks on critical infrastructure.

If the first cybersecurity paradigm was focused on keeping adversaries out, building firewalls and digital moats, the next evolution must move us towards resilience with a more holistic approach where machine learning and other innovative technology empowers teams and where policies protect humans from themselves. Ironically, in this paradigm humans are empowered and no longer the weakest link in the chain, but the supervisory layer to integrity.