Chapter 10
Digital Twin Industrial Immune System: AI-driven Cybersecurity for Critical Infrastructures

Michael Mylrea, Matt Nielsen, Justin John, and Masoud Abbaszadeh

Abstract Innovative advances in machine learning (ML) and artificial intelligence (AI)-driven cyber-physical anomaly detection will help to improve the security, reliability and resilience of the United States' power grid. These advances are timely as sophisticated cyber adversaries are increasingly deploying innovative tactics, techniques and technology to attack critical energy infrastructures. Defenders of these modern infrastructures need to better understand how to combine innovative technology in a way that enables their teams to detect, protect, respond and endure attacks from complex, nonlinear and rapidly evolving cyber threats. This chapter (i) explores how AI is being combined with advances in physics to develop a next-generation industrial immune system to defend against sophisticated cyber-physical attacks to critical infrastructure; (ii) provides an overview of the technology and explores its applicability to address the needs of cyber defenders to critical energy infrastructures; applicability is explored through opportunities and challenges related to human–machine teams as well as the process and technology; (iii) includes validation and verification of findings when the technology was tested defending against stealthy attacks on the world's largest gas turbines; (iv) explores how the AI algorithms are being developed to provide cyber defenders with improved cyber situation awareness to rapidly detect, locate and neutralize the threat; and (v) concludes with future research to overcome human–machine challenges with neutralizing threats from all hazards.

M. Mylrea (✉)
Distinguished Engineer, Cybersecurity & Digital Innovation, National Resilience, Inc, Washington, DC, USA
e-mail: michael.mylrea@resilience.com

M. Nielsen · J. John · M. Abbaszadeh
Cybersecurity, R&D for Operational Technology, General Electric (GE) Global Research, Washington, DC, USA

W. F. Lawless et al. (eds.), *Systems Engineering and Artificial Intelligence*,
https://doi.org/10.1007/978-3-030-77283-3_10

10.1 Introduction

A digital transformation of critical energy infrastructures is underway that is rapidly digitizing, networking and automating the energy value chain. Today's smart energy systems unlock new value in modernizing the grid that is increasingly interoperable, two-way, agile and flexible in incorporating distributed energy resources. While grid modernization helped transition energy usage and consumption to lower carbon, sustainable, renewable energy, new cyber-physical security challenges in securing critical energy delivery systems and associated operational technology (OT) have accompanied this digital transformation. The rapid digital transformation of our critical systems has significantly increased its attack surfaces by combining cyber-physical systems, software and hardware, information technology (IT) and operational technology (OT). This has created new challenges to identify, monitor and protect these critical systems. Improvements are needed for real-time cyber-physical situational awareness and monitoring the cyber threat-attack surface in terms of control systems, automation and other operational technology.

10.1.1 Overview

While cybersecurity technology continues to improve, the attack surfaces of the power grid have expanded significantly, leaving a number of major cyber gaps remaining. For one, most cyber defenses and monitoring solutions are ineffective in detecting sophisticated attacks targeting operational technology, such as energy delivery and industrial control sytems. Zero-day exploits, insider and supply chain attacks continue to evade and defeat cyber defenses and intrusion detection systems. These systems originated from securing information technology across a business enterprise and defending against known malware, malicious packets and other attacks that are easy to catalogue in a library as signature herusitics. However, OT found in various energy delivery systems, such as electricity infrastructures presents new challenges as the protocol, malware signatures, and tactics, techniques and procedures used by adversaries also differ significantly. Moreover, a number of more than 3,000 energy utilities in the U.S. lack basic cybersecurity defenses to identify and monitor their critical cyber OT assets. Thus, the detection of sophisticated adversaries is limited—usually too late or reactive, only after the damage has been done—enabling them to perist their malicious activities in critical systems and networks and often without being detected.

To overcome these limitations, solutions must advance from security to resilience and provide more holistic cover for critical OT in electricity infrastructures. To realize these goals, GE research has been working closely with the U.S. Department of Energy, DARPA and our energy utlitiy partners to leverage advances in artificial intelligence (AI) and machine learning (ML) to develop an industrial immune system for critical operational technology, from wind turbines to combined-cycle power plants, and from hydropower to oil and gas pipelines. In one of the world's first demonstrations of AI/ML self-healing neutralization at scale and with accuracy (99%), GE Research demonstrated the ability to neutralize sophisticated cyberattacks on the world's largest gas turbines. While this scientific accomplishment highlighted advances in AI/ML cyber-physical anomaly detection, a number of challenges remain. Overcoming these challenges requires scientific advances and research that combine complex problem sets at the nexus of people to accomplish these goals, process and technology to secure high assurance systems that are increasingly autonomous. Some of these areas explored in this chapter include explainable AI (XAI), invariant learning and humble AI. These advances are critical to improve the data fusion, trustworthiness and accuracy of AI-driven technology and its application in empowering human–machine teams.

Additional advances are needed not only to detect and challenge decision support for complex autonomous systems but also to the system designers and operators who do not understand and/or trust the decisions that the algorithms are making. This lack of explanation, context and trustworthiness in the algorithms slows adoption and impedes innovation. End users are hesitant to trust the algorithms because they cannot correlate AI-driven machine decisions with the physics and their own domain of experience. In a safety critical system, not understanding the physics and how algorithms are reaching their decisions curtails innovation in next-generation system design and deployment. Overcoming these barriers would help owners, operators and other complex systems stakeholders better understand how algorithms are learning and making decisions, allowing the translation of big data sets into actionable intelligence. Advancements in explainable AI (XAI) would remove these barriers to innovation and provide significant value in advancing the science of sense-making, context and trustworthiness of AI systems.

10.1.2 Cybersecurity Technology Gaps for Advanced Detection, Protection and Monitoring Solutions

Grid modernization has spurred the integration of distributed energy resources (DER's) and the electricity infrastructure that is increasingly digitized, networked, automated and complex in its communications using multiple languages and protocols between an increasing number of parties (Qi et al., 2016). 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. Fig. 10.1 highlights how grid cyber defenders have responded to the cyber threats posed to (DERs) with various cybersecurity solutions that try to segment and

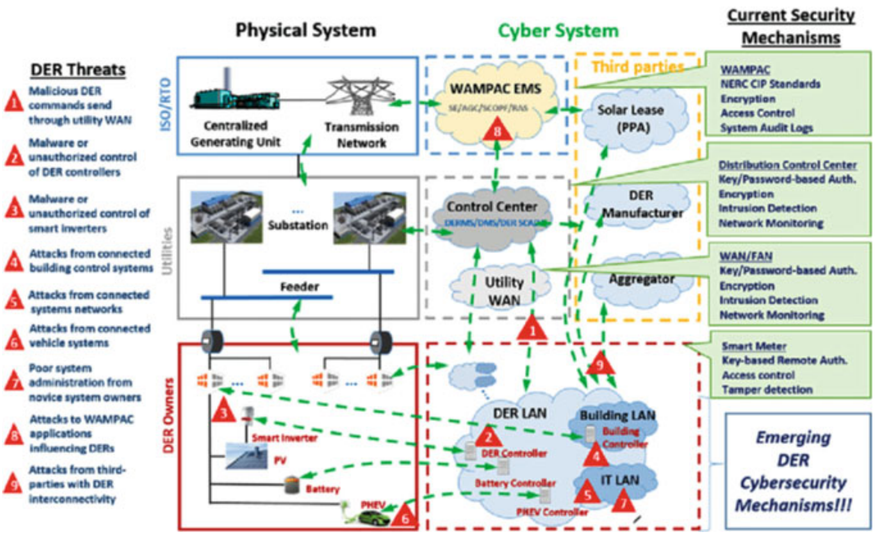

Fig. 10.1 Cyber-physical threat to distributed energy resources (DERs)

to provide "air gaps" for critical systems. However, these cybersecurity solutions do not provide cyber resilience against sophisticated threat actors nor hybrid cyber-physical events (e.g., extreme weather, insider threats, human error, supply chain attacks on software, hardware, etc.) (Fig. 10.1).[1]

As a result, there are numerous cybersecurity gaps for the advanced detection, protection and monitoring of energy delivery systems, networks and interconnected energy delivery systems. These gaps could potentially be exploited to cause the degradation of service and potential cascading failures to the power grid. However, due to the many gaps existing in detection and monitoring, it is difficult to quantify the threat and risk. Increased monitoring and detection of electricity infrastructure may give the perception that attacks to the grid are increasing when in fact this increase is a measure of an improved cyber situation awareness. When an industry article and/or publication suggests there is an increase in cyber-attacks on the grid, is that because monitoring and detection technology have improved, or because threat groups are increasingly targeting the grid? Currently, there is a major gap in the research and data available to quantify these risks. This gap makes it difficult for energy utilities to make strategic investments to buy down the risks to them that are greatest based on the threat. Another major cybersecurity gap for advanced detection, protection and monitoring is found with the increasing penetration of distributed energy resources (DERs, Greenberg et al., 2018; also, see Utility Dive).[2,3] Increased connectivity and

[1] Qi et al. (2016).

[2] https://www.utilitydive.com/news/security-and-distributed-resources-an-attacker-will-eventually-get-in-s/565966/.

[3] Greenberg et al. (2018).

the 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 (Lee, 2013).[4] Any holistic solution requires a comprehensive approach of human and machine, or people, process and technology. But many other gaps remain.

Policy Gaps—Currently the North American Electric Reliability Corporation Critical Infrastructure Protection (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 grid infrastructures are vulnerable to sophisticated cyber-attacks.

Technology Gaps—The data and connectivity requirements needed to improve grid edge and DER management—increased awareness, controls direct-level electrical loads, manage capacity constraints and reverse power flows—has significantly expanded the attack surfaces of our nation's grid. For example, solar energy systems grid-support functions can be manipulated to diminish reliability and damage electricity infrastructure. Securing photovoltaic (PV) system critical communications at rest as well as in transit to aggregators (residential, utility, commercial), utilities and other grid operators is increasingly challenging due to increased internet connectivity and digitization (Johnson, 2017)[5] as well as communication protocols that prioritize interoperability but lack basic encryption and authentication mechanism (Onunkwo et al., 2018).[6]

Together, current policies, processes and technologies prioritize interoperability and connectivity but they do not provide the high fidelity cyber situational awareness needed 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 flows; supply chain attacks can push updates to be behind the metered systems to add or drop load in a way that could potentially cause a grid level event; insider attacks can cause instabilities like sub-synchronous resonances, and man in the middle attacks can amplify weak grid conditions, just to name a few.

10.1.3 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 solutions 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

[4] Lee (2013).

[5] Johnson (2017).

[6] Onunkwo et al. (2018).

recover from cyber-physical threats and vulnerabilities targeting operational technology (OT). If an adversary attacks, manipulates or compromises a critical energy delivery system, GE's Digital Ghost helps to detect anomalous behavior, locate and neutralize the attack while maintaining the 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, anomalies and associated vulnerabilities and to quantify them; ***Localize***, Isolate and Protect critical control systems and OT (sensors/actuators/drives/controllers); and ***Neutralize*** to autonomously ***Respond*** and ***Recover***, mitigating advanced threats. The ability to review the control logic and to autonomously maintain operations without losing the availability of critical systems is a potential game changer to provide cyber-physical resilience, but many challenges remain.

Cyber defense of critical infrastructure continues to evolve, but cyber adversaries often have the upper hand as their offensive tools improve and the attack surface available to them expands. Cyber challenges remain for 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 improves, 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 the move toward automation. As we digitize, automate and connect systems in critical infrastructure to the internet, this also expands 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 that we can identify, detect, respond and recover to cyber threats and vulnerabilities in subsecond times. Cyber resilience includes not only a hardened perimeter, but it also neutralizes sophisticated attacks once they have been found.

Advances of innovative threat mitigation solutions help to move the industry toward cyber resilience. However, the design and implementation of these advances, such as machine learning algorithms, requires the distillation of large data sets to be intelligently fused with operations. The form of the 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 the next-generation AI cyber-defense technology to detect, localize and neutralize threats in a more effective and autonomous way. To realize this goal requires **the leveraging of 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 further in Fig. 10.2.

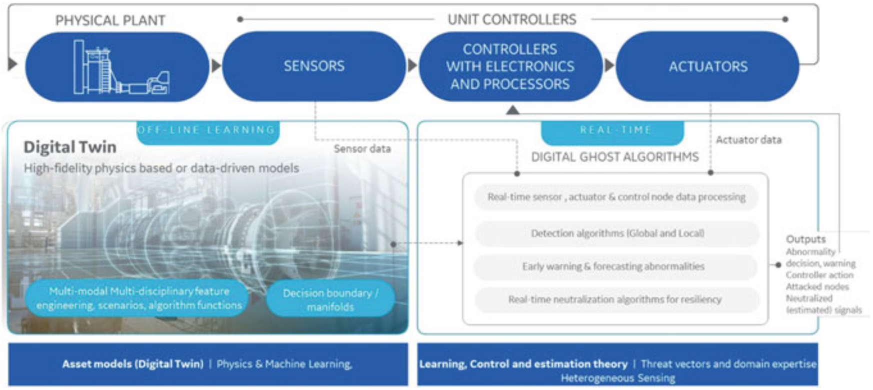
PHYSICAL PLANT
UNIT CONTROLLERS
SENSORS
CONTROLLERS WITH ELECTRONICS AND PROCESSORS
ACTUATORS
OFF-LINE LEARNING
Digital Twin
High-fidelity physics based or data-driven models
Sensor data
Multi-modal Multi-disciplinary feature engineering, scenarios, algorithm functions
Decision boundary / manifolds
REAL-TIME
Actuator data
DIGITAL GHOST ALGORITHMS
Real-time sensor , actuator & control node data processing
Detection algorithms (Global and Local)
Early warning & forecasting abnormalities
Real-time neutralization algorithms for resiliency
Outputs
Abnormality decision, warning
Controller action
Attacked nodes
Neutralized (estimated) signals
Asset models (Digital Twin) | Physics & Machine Learning.
Learning, Control and estimation theory | Threat vectors and domain expertise
Heterogeneous Sensing

Fig. 10.2 Digital Ghost functionality diagram. The example is of a power generation plant. The top portion in the figures depicts a complex system with sensors, 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

10.2 People, Process and Technology Applicability Gap Analysis

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

10.2.1 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 develop the machine learning algorithms, we leveraged our deep domain knowledge of the physics for the systems to establish a matrix of credible cyber-attacks, naturally occurring faults and vulnerabilities in the system. The highest impact abnormalities (i.e., attacks/faults) are chosen for computer model simulations. The high-fidelity Digital Twin models are exercised to define the system’s operating boundaries. Normal operating space is mapped out as well as attack/abnormal operating spaces. The machine learning algorithm developed from these defined scenarios is intended to differentiate between a naturally occurring system fault or a 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. The next step is to establish the decision boundary, called a 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 algorithms on a computer platform connected to the targeted system. Once deployed, the detection algorithm performance is reviewed and continuously monitored.

Technology Gaps—The following four are the areas in technology gaps that need to be closed: (i) Unlike IT solutions which are easy to enumerate and inventory by scanning, operational technology includes a diverse attack surface that is often connected through both internet protocol (IP), serial and other connections. (ii) Proprietary protocols are often vulnerable by design as vendors prioritize functionality, ease of use and cost over security. (iii) Firewalls, network and host intrusion detection systems are limited to defending against malicious signatures, but they are not in their libraries of attack 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) And resource-intensive tuning can be required for AI defense critical solutions to be integrated into existing technology stacks for security information and event management (SIEM).

Process and Policy Gaps: As AI solutions improve attack detections 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 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 do not read these logs. Moreover, additional networks or systems that are connected to these critical cyber assets can provide an attack pathway if they are not secured.

People Gaps: Machine learning algorithms that have high-false positive rates create prohibitive operations and maintenance requirements for security teams. Cybersecurity teams have been traditionally IT-focused; however, the convergence of IT/OT in critical infrastructures has increased the responsibilities and created new workforce development challenges for them. Some innovative new tools require training, but 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 invariant learning and humble AI explored in this chapter highlight how researchers are overcoming these gaps.

10.2.2 Attack Localization

This phase develops a software algorithm that localizes 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 a previous attack through propagation of the effects in the closed-loop feedback control system. This feedback 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 assets and systems, the sub-second time requirements for effective detection and localization are a major gap for most cyber-defense solutions. Moreover, there is a lack of real-time detection and localization solutions to respond to cyber-attacks. Visibility of the data and the 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 the 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. The speed of response for malware and infiltration mitigation to an attack is a critical gap that needs to be met to maintain reliable, safe and secure plant operations. Finally, critical OT is difficult to monitor, especially in converged IT/OT environments that combine various cyber and physical legacy and modern system 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 a loss of power or degradation in output; but not how to localize a cyber-attack. There is a real risk that adversaries could imbed themselves onto 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 like a power plant is no trivial task. In addition, the resource gaps noted for detection and localization have similar and related issues related to localizing an actual system that faulted; this problem is especially true during a transient event or when there is a highly variable stochastic load, events that create a lot of noise and that challenge 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., the loss of power or degradation in output).

10.2.3 Attack Neutralization

Advancing from cybersecurity to cyber resilience requires improvements in neutralization and the ability to recover and endure all hazards, ranging from sophisticated cyber-attacks to naturally occurring events. Neutralization also requires the ability to remove the effects of attacks on the monitoring nodes so that the system can continue to function even in the presence of attacks. It uses the observability provided by the deemed trusted nodes (non-attacked nodes) in its calculations. It will enable true operational signals to be provided to the control system on a continuous basis while informing the operator when attacks are detected.[7] If an attacked node lacks observability, then the error in its replacement estimate may be unacceptably large, preventing continued operation.

Further research into how to autonomously identify critical nodes with poor observability is required to advance secure communication application methods for critical OT. Since neutralization cyber solutions will interact directly with the critical control loops of an operating asset, additional research focused on control stability is needed. Operators must be able to trust that the combined system (Digital Ghost, controller and asset) will behave with stable operations. Operating regions may exist where stability cannot be guaranteed, e.g., outside of the boundaries used for training. The asset's allowable operations must be limited, and research into autonomously identifying these restricted regions is required.

Technology Gaps: For critical OT assets and systems, there are sub-second time requirements for effective communications. Sub-second time requirements are demanded by the dynamics of the system. For many of the critical assets we want to be monitored with Digital Ghost, we need to be able to attack the nodes estimated via the neutralization module at the timescales required by the dynamics of the system. Cyber resilience requires the ability to both detect and localize rapidly to effectively and accurately neutralize an attack or anomaly. Sophisticated cyber-attacks, zero-day exploits, hybrid cyber-physical attacks, 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, but not security. This gap creates additional challenges related to neutralization. The TRISIS cyber exploit was exemplary of these design vulnerabilities where a safety instrumented control system was exploited in a sophisticated attack on operational technology.

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

People Gaps: (i) Lastly, trust between Digital Ghost's neutralization algorithm and the operator must be established. During a cyber-attack, the operator must be presented with clear, concise and understandable information to quickly ascertain

[7] John et al. (2020).

the context and impact to the operations of the machine protected by Digital Ghost. Neutralization leverages concepts from AI/ML, while operators and control engineers often prefer “deterministic” algorithms governing control logic. Research into the more effective autonomous system–human operations is required for neutralization to be an accepted mitigation approach.

10.2.4 Man Versus Machine Anomaly Forecasting and Detection

Anomaly forecasts enable the early detection of stealthy attacks which could otherwise remain in an asset for days or months without being caught. It also enables the early engagement of the system’s operator or the automatic accommodation in a cyber incident. Furthermore, the anomaly forecast system can predict future system failures/malfunctions and can 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 damage may have already been done, with an impact too large to be accommodated. These advances provide an early warning capability to attack detection so that a security breach is detected and alarmed at an early stage both for an operator’s response and for an attack accommodation by the system.

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

10.3 Digital Ghost Research Findings and Future Research

10.3.1 Invariant Learning

Measuring both anomalies and invariances in deep networks for a complex system-of-systems like the power grid is not an easy task. For one, the increased penetration of stochastic and intermittent distributed energy resources further complicates the essential pattern recognition tasks to be able to flag anomalies and variances. Recent research shows advances in training deep architectures in a supervised manner to be invariant to the multiple confounding properties and input transformations found in electricity infrastructures (Goodfellow et al., 2009).[8] Future research examining how to enhance invariant machine learning to improve the cyber-attack detection and accommodation (ADA) accuracy of the Digital Twin models that identify and protect against cyber-physical attacks on critical energy systems and infrastructures is essential.

Modeling a complex system-of-systems for an electricity infrastructure is challenging due to the number of issues from bias offsets between the actual values of the key nodes being monitored and those found in simulations to be “noise” in the system. What appears as an anomaly could be caused by human error, computational error, a naturally occurring weather and ambient event, an increase in supply and demand, a cyber-attack, or a hybrid cyber-physical event. Moreover, adversaries could potentially exploit continuous machine learning biases with the next-generation machine learning attacks that slowly bias key nodes such that the continuous system “learns” this incorrect behavior and treats it as normal. To overcome these challenges, the 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 alerts of cyber events that are clearly displayed to the cyber defenders of a grid especially when they are already distracted with many tools, screens and the day-to-day challenges of keeping the grid reliable and balanced.

These findings point toward 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 to alleviate model mismatches. However, continuous learning could also create a potential new cyber-attack surface where an attacker slowly biases key nodes so that the continuous system again “learns” this incorrect behavior and treats it as normal. Advances in invariant learning are needed to mitigate this manipulation of continuous learning algorithms.

10.3.2 Autonomous Defense: Critical Sensors Identification and Trust

Self-healing complex system-of-systems are the holy grail of cybersecurity research and development. Conference organizers highlight the many challenges that affect “the design, performance, networks operating autonomous human–machine teams” (Lawless et al., 2020).[9] Research findings from testing Digital Ghost’s neutralization algorithms suggest that these challenges increase when human teams lack the observability and context for a complex transient system such as a gas turbine. This

[8] Goodfellow et al. (2009).

[9] Lawless et al. (2020).

gap suggests that advances in autonomous cyber defenses must prioritize the 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 to improve the ability of machine learning algorithms to decide which sensors have poor observability before moving to deployment. In a complex, transient system-of-systems, there is a need to improve the observability and trustworthiness of critical energy delivery sensors to autonomously protect, detect, recover and neutralize cyber-physical threats. In absence of these capabilities, that near-terms opportunities to improve the state-of-the-art for neutralization, including determining the sensors that lack observability for neutralization, are needed to create an alert for human operators; this alert would signal the inability for neutralization to provide corrective action if one of these nodes were attacked. Applying advanced encryption and authentication mechanisms for these sensors via trusted platform modules and other solutions is also ripe for future research and exploration. This achievement would help the information security community to better understand how to improve control theory methods that combine with human–machine teams so that machine learning algorithms can empower cyber defenders to better determine the integrity and trustworthiness of critical sensors.

10.3.3 Humble AI

Humble AI is making valuable advances in marrying man and machine, answering such questions as: How can the algorithms alert the operator of a potential decrease in accuracy or confidence in its threat 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 extrapolation happens? If so, should Digital Ghost or other advanced AI cyber-defense halt operations? Or does the system continue but express reduced confidence in its results? The next-generation AI-cyber-physical anomaly detection and neutralization requires the continuous improvement of ML/AI methods that are agile, adaptable and evolve for complex, nonlinear and changing threats. R&D findings from the 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 critical energy delivery systems move away from training and into regions not simulated previously. In the field, if operating conditions move away from training and extend into regions not previously 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.

10.3.4 Explainable AI (XAI)

Explainable AI or XAI is the ability of AI-based machines to explain the reasoning underlying their decisions in a way that is understandable to humans. Many challenging questions and/or gaps remain, such as: how do we develop intuitive, trustworthy explanations of how and why our AI algorithms arrive at decisions? How do we do this in a way that is easy to interpret, visualize and use to empower human–machine teams?

How do we trust the black-box nature of deep neural networks? That is, numerous parameters in deep neural networks (DNNs) add complexity that is hard to interpret and explain. As a result, algorithms and models can learn and misinterpret representations from the data differently than humans. This creates issues with trust, ethics and biases.

Answering these questions will help to improve the state-of-the-art of (AI/ML) algorithms with a focus on advancing XAI physics-based anomaly detection in complex systems.

If successful, human–machine teams will be able to both trust and understand *how* the ML/AI algorithms arrive at their solutions. This collaboration can be done through advanced human–machine interfaces containing easy-to-understand visualization techniques. This result is essential for machines to be trusted in making autonomous/semi-autonomous decisions, especially for kinetic platforms that are increasingly autonomous as well as for safety and other mission-critical applications that determine diagnostics and cyber-physical security.

Advances in AI require both human operators and machines to understand and trust how the ML/AI algorithms are arriving, or are unable to arrive, at their solutions via human–machine interfaces and intuitive visualizations.

Machine-learning-based approaches of anomaly detection often result in a classification decision along with an anomaly score. However, the contribution, ranking and significance of each of the input variables/features, the causality directions, the effect of the size of the training data set and the reasoning path in the algorithm leading to a particular decision are often obscure. An example of this is shown in Fig. 10.3, where several signal traces are presented as well as features used as inputs for an AI/ML-based, anomaly-classification system, i.e., Digital Ghost. These traces

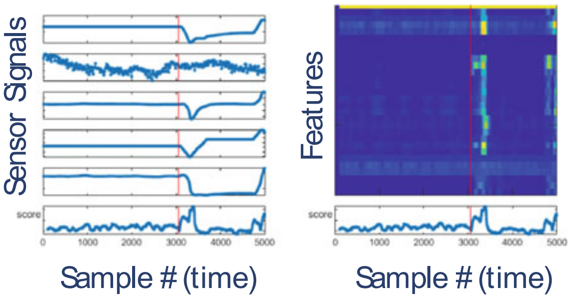

Fig. 10.3 Sensor signal traces and features captured during GE's recent cybersecurity demonstration using an operating heavy-duty gas turbine. The red line indicates the time at which synthetic cyber-attack was injected

were captured on GE Power's test facility during a demonstration wherein sophisticated synthetic cyberattacks were injected into a critical network of an operating gas turbine, and Digital Ghost correctly detected and identified the impacted gas-turbine sensors. While novel in improving the applied science of AI cyber-physical anomaly detection, it has proved difficult to explain to human operators how the complex algorithms arrived at the correct inferences (attack and attacked node) in a manner that an operator can understand, trust and then act upon.

Future research on explainable XAI will advance AI/ML capabilities without reducing the fidelity and accuracy of the detection, localization and neutralization capabilities. It is essential that Digital Ghost's next-generation cyber-physical, anomaly-detection and neutralization algorithms reduce their technical complexity and that they are intuitive to grid operators and cyber defenders. This explanation creates a number of human and cyber-physical integration challenges that could be explored with future research on how best to integrate humans and machines. Lessons learned from (DG) research have helped to develop complex algorithms, some of which are constructed using machine learning and AI techniques. Future operators, however, may still be skeptical because of the complexity and non-intuitiveness contained within the highly nonlinear algorithms of Digital Ghost.

10.4 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 in this chapter highlighted how this digital transformation must marry man and machine. Similarly, research findings also suggest that human–machine teams can be empowered but also blindsided by AI by being 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 a responsibility to ensure more holistic cyber resilient policies, processes and technology.

Realizing this goal is imperative as the U.S. electricity infrastructure will require a holistic approach of people, policies and technology. Research findings suggest the successful adoption of next-generation technology, such as the AI algorithms found in Digital Ghost. Findings also suggest that innovation should not happen 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 the cyber defenders of our nation's power grid. This result is especially true as sophisticated cyber adversaries are increasingly deploying technology such as AI combined with stealthy tactics and techniques to attack critical energy infrastructures. 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, nonlinear 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 from the loop. This is especially true when defending against the diverse, complex, nonlinear and rapidly evolving threats of human adversaries executing sophisticated cyber-physical attacks on critical infrastructures.

If the first cybersecurity paradigm was focused on keeping adversaries out, building firewalls and digital moats, the next evolution must move us toward 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 that provides integrity.

References

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).

Greenberg, A., Dreyfuss, E., Barrett, B., Gold, D., Lapowsky, I., & Newman, L. H. (2018). How hacked water heaters could trigger mass blackouts. Wired, August 2018.

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.).

Johnson, J. (2017). Roadmap for photovoltaic cyber security. Sandia Technical. Report, SAND2017–13262.

Onunkwo, I., Wright, B., Cordeiro, P., Jacobs, N., Lai, C., Johnson, J., & Hutchins, T. et al. (2018). *Cybersecurity assessments on emulated DER communication networks.* Sandia Technical Report.

Lawless, W. F., Mittu, R., Sofge, D., Shortell, T., & McDermott, T. (2020). AI welcomes systems engineering: *Towards the science of interdependence for autonomous human-machine teams.* AAAI Spring Symposium 2020 Abstract.

Lee, A. (2013). Electric sector failure scenarios and impact analyses. *National electric sector cybersecurity organization resource (NESCOR) technical working group 1.*

Qi, J., Hahn, A., Lu, X., Wang, J., & Liu, C.-C. (2016). Cybersecurity for distributed energy resources and smart inverters. *IET Cyber-Physical Systems: Theory & Applications., 1*, 28–39. https://doi.org/10.1049/iet-cps.2016.0018