HADMÉRNÖK
VÉDELEMINFORMATIKA
21. évfolyam (2026) 1. szám 5–24.
DOI: 10.32567/hm.2026.1.1

Dániel Griffiths[1]

Analysing Cognitive-Computing-Based Countermeasures to Cyber Threats against Critical Infrastructures

Abstract

Critical infrastructures are vital to the functioning of modern societies but are increasingly exposed to sophisticated cyber threats due to heightened convergence of information technology and operational technology environments. The inherent complexity and presence of legacy components within these systems significantly expand the attack surface, making traditional security measures insufficient. This work analyses the evolving threat landscape facing critical infrastructures and presents a comprehensive assessment of cognitive computing-based countermeasures. Emphasising recent advances in cognitive computing, the article explores adaptive defence mechanisms capable of evolving alongside emerging threats. The analysis highlights the effectiveness and limitations of these approaches, their applicability within cyber-physical systems, and their potential to bridge gaps between IT and OT security. Ultimately, the research underscores the importance of integrating these techniques into cybersecurity strategies to enhance the resilience of critical infrastructure systems.

Keywords: cognitive computing, machine learning, cybersecurity, critical infrastructure, cyber-physical systems

The attack surface of critical infrastructures

Critical infrastructures play a vital role in the functioning of modern societies, ensuring the smooth operation of everyday life. These infrastructures are generally defined as the totality of physical and virtual assets, systems and services whose failure, destruction, or prolonged disruption could have a serious, widespread impact, threatening

[1] PhD student, Óbuda University, Doctoral School on Safety and Security Sciences.

the safety of the population, the functioning of the state and the stability of the economy. As a result, attacks against critical infrastructures may have profound societal, economic and national security implications. A successful attack may disrupt daily life, economic activity or even public order. Furthermore, the targeting of strategic services such as energy, communications, or transport systems can impede governmental functioning and threaten national sovereignty.[2]

Modern critical infrastructures are highly interconnected and interdependent, so much so that an attack on a single component may trigger cascading effects across multiple systems.[3] Advanced attacks against cyber-physical systems can cause long-term damage, undermine system reliability, and demand substantial financial resources for recovery.[4] High-profile incidents may also generate social panic and erode public confidence in institutions, thus contributing to long-term societal destabilisation.[5] These risks necessitate the prioritisation of critical infrastructure within national cybersecurity strategies, yet effective defence is hindered by the complexity and limited transparency of such systems.

The ongoing digitisation of industrial environments has further intensified these challenges. The cyber-physical systems typically found in critical infrastructures differ significantly from traditional IT systems and operate under very different constraints. These difficulties are compounded by the legacy nature of many critical infrastructures, a significant portion of which was constructed in the mid-20th century or even earlier.[6] Such systems were originally designed as closed, physically isolated environments, with security based primarily on obscurity rather than true cyber resilience.[7] Increased connectivity has since exposed these systems to modern threats, while industrial control system (hereinafter ICS) manufacturers historically placed limited emphasis on cybersecurity, relying on proprietary protocols and physical isolation as protective measures.[8] Consequently, many known vulnerabilities remain unpatched due to unsupported operating systems, long lifecycles and stringent operational requirements.[9] Continuous operation and high availability constraints further limit opportunities for modernisation, patching and security testing, given the severe societal and economic consequences of service interruptions.[10] These technical challenges are exacerbated by the relative specialisation of cyber-physical system security and a persistent shortage of skilled cybersecurity professionals, which disproportionately affects the resilience of critical infrastructures.[11]

A central challenge in defending critical infrastructures is the exposure created by external attack surfaces, which arise primarily at interfaces connecting infrastructure components to public or less controlled networks. In many cases, such systems are

[2] Riggs et al. 2023; Lis–Mendel 2019; Haig–Kovács 2012.
[3] Palleti et al. 2021; Wang et al. 2018.
[4] Baezner–Robin 2017; Jermalavičius et al. 2025.
[5] Tabansky 2011.
[6] Osei-Kyei et al. 2021.
[7] Assenza et al. 2020.
[8] Hurst–Shone 2024; Daricili–Çelik 2022.
[9] Wilson 2014.
[10] Booth et al. 2019; Assenza et al. 2020.
[11] ISC2 2024; ISACA 2024.

inadvertently exposed due to misconfiguration or inadequate network segmentation.[12] Industrial control systems are especially vulnerable, as internet-connected devices can be rapidly identified using automated search engines such as Shodan, enabling targeted attacks.[13] These systems frequently lack robust authentication and encryption, thus creating trivially simple opportunities for unauthorised access, traffic interception or manipulation.[14]

Attackers commonly exploit vulnerable services and open ports associated with outdated, improperly configured, or unpatched software, using automated discovery and exploitation tools. Well-documented attacks against critical cyber-physical systems, such as Stuxnet,[15] Industroyer[16] and Triton,[17] demonstrate how adversaries can gain initial access through exposed interfaces or insufficiently protected remote access and subsequently perform data theft, system manipulation, or physical sabotage. The rapid proliferation of industrial Internet of Things devices further expands the external attack surface, as direct connectivity to cloud-based management platforms introduces additional attack vectors when connections are inadequately secured.[18]

A further major source of risk arises from the exploitation of trusted relationships. Rather than targeting critical infrastructures directly, adversaries may compromise less protected partners, suppliers or service providers from the target's supply chain.[19] These attacks exploit established trust relationships based on routine data exchange, software updates, remote management, or certificate-based authentication. Software supply chain attacks are particularly dangerous where critical infrastructures rely on widely used software or third-party components that are regularly updated. By compromising a software developer or distributor, attackers can inject malicious code into legitimate update mechanisms, which are then automatically deployed by operators who trust the source. Such attacks are difficult to detect, as they abuse valid digital signatures and legitimate distribution channels.

Trusted relationships also arise from interoperability and permissive access arrangements between interconnected systems.[20] Critical infrastructures often operate multiple integrated networks with shared connectivity or authentication mechanisms, enabling attackers who compromise a less secure subsystem to pivot into more sensitive environments. These trust-based connections can produce cascading breaches, allowing a single compromise to undermine multiple layers of defence. The risks associated with this are further exacerbated by increasing reliance on third-party services such as remote maintenance and cloud-based monitoring.[21]

Cyberattacks against critical infrastructure are not limited to technical exploitation, but frequently rely on human manipulation or physical intrusion. Despite the

12 Leszczyna–Egozcue 2013.
13 Wilson 2014.
14 Leszczyna–Egozcue 2013.
15 MITRE Corporation 2025.
16 MITRE Corporation 2024a.
17 MITRE Corporation 2024b.
18 Hurst–Shone 2024.
19 Hurst–Shone 2024.
20 Leszczyna–Egozcue 2013.
21 Hurst–Shone 2024.

deployment of technical safeguards, the human element remains a critical vulnerability.[22] Social engineering attacks exploit such weaknesses by manipulating individuals into disclosing sensitive information, granting unauthorised access or performing actions that undermine operational security. Attackers may impersonate trusted figures such as IT personnel, managers, or external partners, subsequently coercing users into compromising their credentials or systems and enabling undetected access to sensitive resources.

Insider threats present a related risk, where individuals with legitimate access may intentionally or unintentionally facilitate breaches due to negligence, coercion, financial incentives or personal grievances. Such actors can misuse authorised privileges to exfiltrate data, sabotage systems or assist external adversaries. These threats are particularly concerning in critical infrastructure environments, given their strategic importance and the potential involvement of foreign actors seeking to undermine national security. Insider attacks are difficult to mitigate because they exploit trust-based organisational structures and access controls that are essential for normal operations.

Physical breaches constitute a distinct but closely related threat vector. Direct physical attacks aim to disrupt or destroy infrastructure components through forceful means, producing immediate and severe consequences that require significant recovery efforts. More sophisticated cyber-physical attacks leverage physical access to enable digital compromise rather than outright destruction.[23] Attackers may gain entry under the guise of maintenance or contracting activities and introduce infected devices into restricted or air-gapped networks. By bridging physical and digital domains, such hybrid operations circumvent conventional cybersecurity measures based on network isolation or remote-access controls, and their physical appearance may obscure the underlying objective of covert digital infiltration.

Research goals and methodology

Taking into account the threat landscape detailed earlier, the primary objective of this paper is to explore the applicability and potential effectiveness of cognitive computing solutions for the protection of cyber-physical systems present in critical infrastructures. One area of focus is the comparative assessment of the cybersecurity maturity of critical infrastructures with traditional IT environments. Building on these differences, one aim of the study is to analyse the extent to which cognitive defensive solutions applied in traditional IT environments can be adapted to critical infrastructure environments and how they could potentially address the previously identified attack surfaces. The research also aims to identify relevant cognitive approaches that are conceptually suitable for responding to complex threats to cyber-physical systems and provide a scientific basis for further targeted research directions.

[22] Ghafir et al. 2018.
[23] Loukas 2015.

The research methodology is built around a literature review approach, with the objective of assessing cognitive computing approaches in the context of critical infrastructure protection. This involved identifying the relevant research areas, along which a structured keyword-based search strategy could be employed. The review draws on peer-reviewed academic publications as well as industry reports to ensure that the analysis covers both theoretical and practical aspects. The search process focused primarily on the most recent and relevant work, which was complemented by the historical sources necessary to understand the conceptual background. The analysis focuses on approaches that are empirically grounded and applicable to cyber-physical environments. The review thematically categorises these approaches and performs a comparative analysis, allowing for the systematic identification of research challenges, current practices, as well as perceived technological and conceptual gaps.

Existing cognitive-computing-based countermeasures

Cognitive computing represents a new generation of information systems that can simulate aspects of human thinking using approaches such as artificial intelligence (hereinafter AI) and the closely related field of machine learning (hereinafter ML). These systems do not simply follow pre-programmed logic, but learn from their environment and make real-time decisions based on information learned from complex data sets.

Over the past decade, these approaches have undergone significant development, driven by digitisation, automation, and the recognition of the economic and innovation potential of generative AI. Cognitive systems are gradually emerging in IT security, enabling more sophisticated defensive measures and supporting the automation of complex tasks. Although their use in protecting cyber-physical systems in critical infrastructures is still in its infancy, cognitive computing offers particularly promising opportunities, and a number of studies have begun investigating how these systems can help address the specific types of threats targeting these environments. The following sections discuss the research areas that have begun investigating the potential use cases for this technology in the field of critical infrastructures.

Network traffic analysis

Traditional signature-based systems use known attack patterns (signatures) to identify threats to cyber-physical systems (hereinafter CPS), meaning they can only detect predefined threats and are thus ineffective against unknown (zero-day) attacks. In contrast, machine learning-based anomaly detection models the normal behaviour of the system and identifies deviations from it as potential attacks. Their advantage is that they do not require prior signatures; are able to adapt to a variety of CPS systems; and can operate with unlabelled data in unsupervised learning, which is particularly useful in volatile and complex CI environments. Over the past decade, there has been a significant increase in the number of research projects dealing with

ML-based anomaly detection for CPSs and CI. Since the early 2010s, there has been a steady increase in the number of publications, mainly in the fields of computer science and engineering, reflecting the multidisciplinary security challenges of CPS/CI systems and the growing importance of cognitive approaches.[24]

There are numerous cognitive computing approaches in the field of network traffic analysis and intrusion detection, all of which essentially rely on some form of machine learning technique. The chosen learning paradigm significantly influences detection capabilities, especially when dealing with challenges such as zero-day attacks, data drift, and issues of scalability and generalisability.

One such research project investigated the detection of malicious CPS network traffic using a range of classical machine learning and deep learning approaches.[25] The tests involved various classification tasks, where k-nn and decision tree models were tested alongside deep learning approaches such as CNN-LSTM architectures. Based on the researchers' results, these approaches provide virtually error-free predictions on the data sets used. The data set chosen by the authors did not contain any "exploit-like" attacks, thus protection against zero-day attacks is questionable and questions remain about the robustness and generalisability of the resulting systems.

A different research article underpins the idea that deep learning is highly effective for intrusion detection in CPS environments.[26] The presented reinforcement learning model demonstrated strong performance on two industrial IoT datasets, particularly in detecting attacks such as port scans and false command injections. Although the study achieved excellent accuracy, the question still remains regarding the types and number of attacks present in the training datasets, as well as the fact that the study did not include a confusion matrix for each attack category, which makes it difficult to determine how robust the protection is (e.g. in the case of zero-day attacks).

There has also been research exploring the possibilities of unsupervised learning, focusing on the problem of zero-day attack detection. This approach uses machine learning algorithms trained only on data representing normal system functionality. These models are then used to identify deviations from this learned baseline. Such systems appear promising with one paper finding the model capable of identifying 23 out of 26 attack scenarios in a real-world accurate dataset.[27] Similar approaches have been proposed using semi-supervised anomaly detection mechanisms that deviate from the earlier previously mentioned unsupervised models by being also trained with a small amount of labelled data. This method also showed positive results in handling zero-day threats, with the added benefits of low false positive rates.[28]

In addition to the research described above, there are numerous other promising research results that suggest that we are talking about a mature field of research that is already working on ironing out late-stage problems.[29] The main open issue left is generalisability and adaptability to zero-day attacks. This could primarily be achieved

[24] Rouzbahani et al. 2020.
[25] Alkahtani–Aldhyani 2022.
[26] Mesadieu et al. 2024.
[27] Pinto et al. 2024.
[28] Vávra et al. 2021.
[29] Pinto et al. 2023.

by producing more comprehensive data sets that place greater emphasis on zero-day-like attacks and exploits and better represent the real network traffic experienced in critical infrastructures. However, the production of such public datasets that reflect real-world environments may continue to raise significant national security concerns and therefore may remain limited.

Research has been published recommending the use of large language models (hereinafter LLMs) to detect network anomalies in critical infrastructure.[30] These may seem particularly promising in less sensitive IT environments, but in systems where the security of critical infrastructure is at stake, it is essential that security teams have complete sovereignty over both their data and the models into which they feed it. With API-based integrations, there is a risk that data of national security importance could fall into the wrong hands, either directly or indirectly (e.g. through third-party cloud-based systems). Even in case of locally run LLMs, the "black box" nature of the models raises serious concerns, particularly with regard to the exact data used to train them. It cannot be ruled out that nation state actors have deliberately introduced poisoned data sets into publicly available training data, which could result in the model's behaviour being not only unpredictable but also deliberately manipulated.

Automated threat hunting

Another promising research direction using cognitive computing is automated threat hunting. One such research article presented a threat hunting architecture that integrates machine learning methods and visualisation techniques to protect critical infrastructures.[31] The system they propose collects and normalises information from SIEMs, log files, PCAP data and OSINT sources to identify suspicious behaviours.

At the core of the system is an anomaly detection mechanism that uses various machine learning algorithms to identify deviations from previously observed normal behaviour in the infrastructure. Methods such as clustering techniques, decision-tree-based approaches and various neural network architectures work together to achieve this goal. These algorithms are capable of highlighting unusual events from log files or network traffic, recognising multi-step, hidden attack patterns, and learning from new data and threat hunter feedback to increasingly accurately distinguish between benign and malicious behaviour. Natural language processing modules enable the system to extract contextual and semantic information from unstructured sources such as analyst reports, incident descriptions, or threat intelligence materials, and combine this with detected activities to further increase detection accuracy. A novel feature of the system is the hypothesis generator module, which uses the results of various machine learning outputs and rule-based filters to generate attack hypotheses and perform probability assessments, enabling automatic prioritisation of incidents. The system is capable of learning responses to recurring, benign events and treating

[30] Markevych–Dawson 2023.
[31] Aragonés et al. 2023.

them with lower priority, thereby bringing new or rare, potentially dangerous behaviours to the forefront.

Although the technical structure and operating mechanism of the system are well thought out, the publication does not detail exactly what types of anomalies the model is capable of detecting, which raises questions about its practical applicability. During validation in a digital twin-based test environment, the authors found that the system was capable of processing large data streams in real time, detecting complex attacks and integrating expert feedback. After six months, the proportion of events correctly classified as harmless increased to 83.49%, while the attack detection rate increased to 86.31%. However, the article does not provide detailed information on the types of attacks detected or the environment in which the system was tested (e.g. the industry or technological background of the infrastructure), which makes it difficult to assess the validity and generalisability of the results.

In another study with a similar goal but a different approach, the authors presented a proactive threat hunting method that uses multiple machine learning algorithms to detect suspicious activities that may occur in critical infrastructures.[32] The outlined approach is more similar to a traditional anomaly detection machine learning algorithm than to the previous true threat hunting-based approach. Similarly to threat hunting in a traditional SIEM system, this approach is akin to using queries that are too noisy to be actual alert rules, but can still reveal genuine malicious activity.

This approach therefore uses a number of machine learning algorithms to detect suspicious activities, which can then be investigated manually by human analysts. The system was tested on open-source bank transaction data to detect attacks against banking systems (mainly targeting SWIFT). Certain classification approaches achieved high accuracy rates of up to 95.7%, but based on the confusion matrices shown in the research, the false negative rate is quite high. It is also important to consider how applicable these results are in practice. A noisy threat hunting query in a SIEM system provides interpretable results, as the underlying logic is transparent, allowing an analyst to continue the investigation. In contrast, with a machine learning-based approach, where the decision logic is completely opaque, an analyst will not know with certainty how to investigate the case further. Therefore, further research is needed on the validity of such approaches, especially in comparison with the hypothesis-based approach outlined earlier.

Behavioural biometrics

As explained at the beginning of this article, compromised users and internal threats pose a serious challenge to the protection of critical infrastructure. A defence mechanism for this can be found in the areas of advanced behaviour analysis, anomaly detection and behavioural biometrics.

[32] Shan–Myeong 2024.

One such approach is a deep learning-based authentication method that analyses users' keystroke dynamics.[33] This approach introduces a novel method in the field of keystroke dynamics behavioural biometrics, which is capable of converting numerical temporal features into visual representations. This allows users' password entry patterns to be evaluated in visual form, enabling the detection of subtle behavioural differences using a Siamese convolutional neural network, which can be used to distinguish the original user from a potential insider attacker.

During the training of the model, it learns to keep the embeddings of samples from the same user close to each other, while placing samples from different users far apart in the multidimensional representation space. A significant advantage of the methodology presented in this research is that it can identify users not only in the case of static, predefined passwords, but also across different passwords and password lengths. The approach was validated using multiple datasets and metrics, and although there are significant fluctuations in accuracy from dataset to dataset (84.0% to 99.2%), the solution demonstrates significant potential for detecting compromised authentication data.

This level of behavioural biometrics and anomaly detection allows us to detect password compromise after a social engineering or phishing attack if the input pattern differs from that of the legitimate user. It is conceivable that with further research, insider threats will also be detectable using similar approaches, as similar non-critical infrastructure-specific research already exists.[34] This would allow the detection of data leaks or the issuance of malicious PLC commands based on keystrokes. A sufficiently sophisticated model could potentially even identify users who have been deceived by social engineering and are acting under duress.

Sensor fusion

Cognitive computing-based methods also play a role in enhancing the physical security of critical infrastructures. One such example has achieved promising results using applied multimodal data fusion and adaptive deep learning approaches.[35] The authors propose an attack detection framework that combines visual surveillance data, Wi-Fi channel state information, and PLC/SCADA sensor data in a machine learning classification framework. This multimodal approach achieves significantly better detection performance than any unimodal approach, as demonstrated by detailed experimental results in the publication.

The test data set was recorded as part of the EU Horizon 2020 STOP-IT project, a research initiative for the protection of critical water infrastructure, in which eight water utility companies participated as consortium members. The dataset covers various data sources, including video streams from security cameras, Wi-Fi signal data, and data from industrial control systems. The ICS sensor data comes from

[33] Budžys et al. 2024.
[34] Wang et al. 2018; Bondada–Bhanu 2014.
[35] Bakalos et al. 2019.

the SCADA systems of real-world water infrastructure and includes the suction and discharge pressure values of two pumps and data on the water level in a water tank. The data was labelled based on predefined scenarios jointly defined by the water utility operators. These labels can be classified into four main categories.

- normal operation
- cybersecurity attacks (targeting ICS systems)
- physical intrusions (including suspicious movements in protected areas, monitored using RGB and thermal cameras)
- cyber-physical attacks consisting of a combination of the above approaches

The solution proposed by the research is a TDL–CNN architecture that is capable of analysing temporal precedents, so that detection results take into account not only current data but also data from hundreds of previous points in time. This reduces false alarms and increases robustness in dynamically changing attack environments where threats can be physical, cyber or a combination. The results are promising, reaching an accuracy of up to 87.62%. Given the high dimensionality and volume of the data, even better results may be achievable by applying newer neural network architectures more suited to multimodal large datasets. In addition, the use of camera data is likely to improve the interpretability of alerts, as this visual information gives security teams an accurate idea of when an attack occurred and what systems were accessed.

Unimodal approaches, primarily using visual data, may also have their place in protecting critical infrastructures. Use cases where visual anomaly detection can identify physical damage may be useful for surveillance and remote monitoring use cases. Although comprehensive research for uses is limited, there have been proposals for using machine learning-based classification algorithms for powerline inspections.[36] The authors in this case only aimed to classify tower types and did not attempt to detect damage or other abnormalities. However, with further development of this approach, the identification of manipulated, damaged towers or foreign objects could also be turned into a visual multi-class classification problem. Further development of the method, for example by adding multimodal sensor fusion outlined earlier (e.g. using Lidar), could further improve sabotage detection capabilities.

Analysis and future research

As discussed in previous sections, the cybersecurity of critical infrastructure has become one of the most important challenges of the digital age. However, this area still does not receive the attention it deserves in research and development involving cognitive computing techniques. While the use of cognitive computing to enhance the cybersecurity of information technology systems has long been an active area of research, research on industrial and operational technologies is much less common. The current lack of prioritisation is not only risky but also irrational, as most IT research could be adapted to critical infrastructure with minor modifications.

[36] Sampedro et al. 2014.

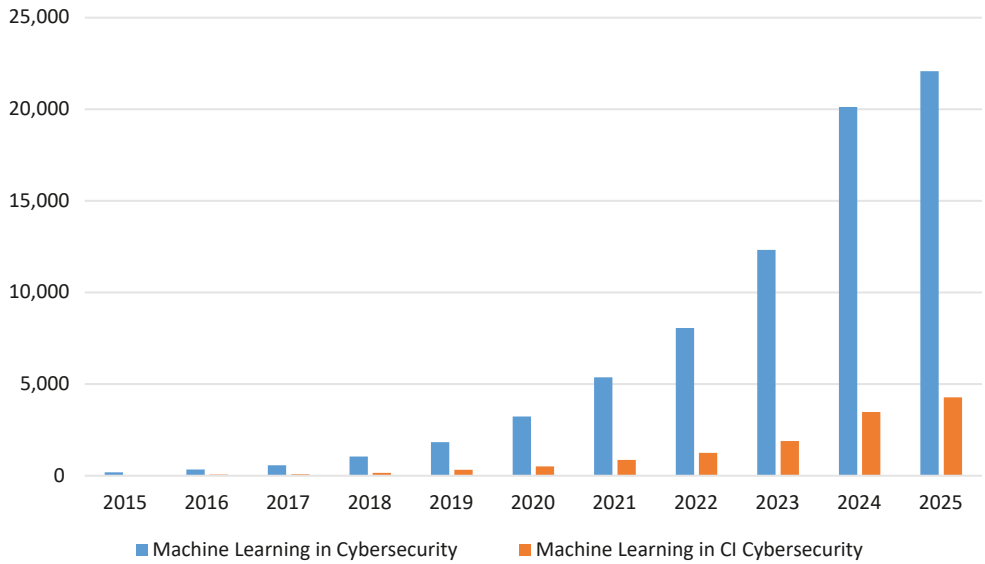

Figure 1: Disparity in machine learning-related Scopus indexed articles
Source: compiled by the author

One of the fundamental reasons why OT cybersecurity does not receive enough attention is the previously established operating paradigm, according to which these systems were physically isolated from IT networks and were closed, proprietary environments optimised primarily for availability as well as secure operation in harsh industrial conditions. As a result, organisations may view OT systems as not requiring the same level of security attention as IT infrastructures. However, this attitude is becoming increasingly untenable.

Technological advances and industry competition are driving ever greater IT/OT convergence, with increased interconnectivity between the two networks. This has made real-time data collection, remote control, and advanced analytics available in systems that previously operated in isolation. IT/OT convergence is further accelerated by the demand for production optimisation, predictive maintenance and centralised management, as well as the proliferation of cloud-based solutions.

As a result, OT systems are moving towards more open, interoperable architectures and integrating more IT-based technologies into their operations.[37] Parallel to the technical and operational convergence of IT and OT systems, it would be logical for security research in both areas to be mutually applicable. Many principles and methodologies, such as access management, network segmentation and incident detection can be adapted from the IT environment to the OT domain, especially in newer, more open systems. Future cybersecurity research should therefore aim to bridge the artificial divide between IT and OT and promote knowledge transfer between the two domains, particularly in the protection of critical infrastructure. Therefore,

[37] Leszczyna–Egozcue 2013.

in the following sections, several approaches using IT-specific cognitive computing techniques are presented and related areas for potential development are explored, which could be applied to the cyber-physical environments of critical infrastructures.

Ensuring network segmentation

The application of machine learning-based network segmentation in information technology networks is a method that already exists and is supported by research.[38] The essence of this approach is that, instead of traditional manual segmentation, large amounts of passively collected IP traffic data is used to automatically form behaviour-based network segments, which can then be separated from each other using firewall or routing rules. IP flow-based network monitoring allows for the collection of metadata from a large volume of traffic at a single central measurement point without direct access to the hosts. This connection-oriented data requires further processing to be transformed into host-oriented behaviour profiles.

During behaviour-based segmentation, features are first selected from the IP flow data that describe the network behaviour of individual hosts. These may include: the number of traffic connections belonging to a host, the number of bytes and packets communicated, the number of peers reached, used ports, protocols and their hourly trends. From these characteristics, a large-scale temporal behaviour pattern is formed for each host.

The authors tested the automated segmentation using several machine learning algorithms, particularly clustering techniques. The research shows that in behaviour-based clustering, the number of actually distinct behavioural clusters is usually much smaller than the number of segments created on an organisational (administrative units) basis, as devices with identical or similar functions often belong to different administrative groups but have similar network behaviour. However, if the subset under investigation exhibits well-distinct behavioural patterns, behaviour-based clustering can show a near-perfect alignment with well-designed logical segments.

Another main application of machine learning segmentation is assigning unknown hosts to existing segments. Here, classification algorithms (primarily clustering algorithms) are used, which assign the host to the most suitable segment based on its behaviour profile. These methods are most accurate (up to 92%) when the behaviour of the segments is well differentiated, while in case of administrative segments, where host behaviour often overlaps, accuracy is around 50–60%.

In principle, this approach could be directly adapted to industrial control technology environments. OT systems are characterised by a narrower set of protocols and traffic that is often more deterministic than in IT networks, which means that behaviour-based profiling could be even more effective. The advantage of the IP flow-based approach is that it does not require the installation of agents on sensitive OT devices, and traffic data can be collected at the edge of the network, thus increasing visibility while maintaining system integrity. Classifying individual OT devices into

[38] Smeriga–Jirsik 2019.

segments and automatically classifying new or unknown devices can significantly contribute to refining segmentation rules while adapting to dynamic changes in the OT network. This reduces the network attack surface and limits lateral movement in the event of a successful attack.

Generative AI for deception

Honeypots based on generative artificial intelligence are a new type of cyber deception tools that leverage the capabilities of LLMs or generative pre-trained transformer models to provide realistic interactive responses to attackers while avoiding the typical limitations of traditional honeypots.[39] These systems are capable of simulating text-based interactions, such as command line terminals or even complete protocols, where the generative model responds to every incoming command or request based on the current context and history of the system. Prompt engineering techniques allow the model's "personality" and context to be precisely defined, such as to mimic a specific operating system, device, or service. To make responses more deterministic, the input prompt and context can be continuously and dynamically edited to ensure the realism and consistency of outputs.

The key to the operation of generative AI honeypots is that when an attacker's command or interaction is received, the system pre-processes the input through a so-called front-end interface layer and then post-processes the response given by the model (e.g. filtering sensitive information or self-disclosing outputs), and finally returns it to the adversary. Context management plays a critical role in ensuring that only command-response pairs that actually modify the state of the system are retained, while the entire command history can be optionally stored for special cases (e.g. history commands). This significantly reduces token usage and increases response coherence, enabling longer, more convincing attacker sequences to be handled without hitting the LLM's token limits.

The use of such honeypots seems feasible in CPS environments, as they could allow simulating the behaviour of industrial control systems, PLCs, DCSs, or even IoT/IIoT devices in the form of textual or protocol-level interactions. By customising the prompt or potentially with some limited model retraining or fine tuning, a generative AI honeypot has the potential to simulate an industrial protocol or interpret and respond to a command set similar to that of a specific PLC. Through its adaptive capabilities, the model could respond to new, previously unseen attack patterns and dynamically expand the simulated behaviour patterns based on attacker interactions, thus avoiding rapid fingerprinting of static honeypots. In addition, this approach minimises the risk of the attacker breaking out of the honeypot, as they cannot execute real code on the backend system.

On the other hand, generative AI honeypots are likely to face a number of technical challenges, especially in OT environments. The determinism of LLMs is only partially guaranteed. Due to the stochastic selection of tokens, different responses

[39] Ragsdale–Boppana 2023.

may be generated for the same command, which could expose deception to a persistent attacker. During long or complex sequences, hallucinations may occur or the token limit can be quickly reached if the entire history of every interaction is always passed to the model. These problems may be solvable by using various prompt and context management and post-processing techniques. Furthermore, the binary protocol-level communication often used by real OT systems can only be simulated to a limited extent with text-based models. Therefore, generative AI honeypots must be combined with other traffic generation tools, or special generative models must be created with unique data sets that are capable of simulating these binary protocols.

Software supply chain monitoring

As discussed in previous sections, supply chain attacks pose a serious threat to industrial control systems, where attackers can cause physical damage or compromise operations by installing malicious or faulty control code. Code analysis could play a key role in preventing and detecting such attacks, as it allows potential vulnerabilities, logical errors, or unwanted functions to be identified before installation, even without running the code.

However, current methods of code analysis for cyber-physical systems, especially in OT/ICS environments, rarely use cognitive computing approaches. Instead, traditional static code analysis techniques are used for formal or semantic analysis of industrial control programmes. These methods typically first translate PLC programmes into an intermediate representation (hereinafter IR) that contains simple instructions such as assignments, jumps and conditional branches.[40] This IR normalises the different PLC languages, making them uniformly analysable. Next, a control flow graph is created from the IR, which depicts the possible execution paths of the programme. Abstract interpretation is then performed on this graph, during which the value ranges of each variable in the programme are assigned to the various possible execution branches. Based on this data, predefined rules and checks are executed to identify syntactic errors, potential runtime errors, unwanted variable values, or even dangerous logical branches. Such static analyses are suitable for detecting when a malicious or otherwise flawed code snippet could cause dangerous behaviour. However, these methods often generate a large number of false positives because they cannot take into account the runtime context, such as whether parts of the code are actually accessible during real operation or how the physical environment limits possible execution paths. Furthermore, such static rule-based approaches are limited or unable to detect unknown weaknesses, such as zero-day vulnerabilities or attack vectors that were not originally considered by the developers. In contrast, machine learning-based approaches are typically more generalisable and can therefore be more effective against unknown attacks.

In other areas, such as classic information technology systems or software development processes, cognitive computing and code analysis approaches are already

[40] Stattelmann et al. 2014; Zonouz et al. 2014; Zhang et al. 2019.

widespread.[41] These systems were typically trained with large amounts of source code, machine code, binaries, configuration files, or even documentation elements and use ML algorithms, including various specialised language processing techniques. Similar to classical approaches, the first step in AI/ML-based code analysis is typically the production of an intermediate representation of the source code, such as an abstract syntax tree, control flow graph or simple tokenisation, followed by feature extraction. The machine learning model then classifies or predictively evaluates the code snippets, for example, by predicting whether a given code snippet contains vulnerabilities, backdoors, anomalies, or other undesirable behaviour.

LLMs offer a new perspective on detecting these software vulnerabilities. Not only are these models capable of understanding the syntactic and semantic structure of source code, but by learning from a huge codebase, they implicitly incorporate various programming patterns and contexts related to previous vulnerabilities.[42] Such an approach enables LLMs to automatically identify and determine which input points in a given software project represent potential threats and which API calls or functions are the locations where improperly sanitised data could cause security vulnerabilities. This not only increases the effectiveness of vulnerability detection, but also significantly reduces false positives, especially when LLMs are able to decide on the relevance of an issue by taking into account the context of the entire project. Furthermore, the use of LLMs offer even greater potential for discovering new or previously unknown vulnerabilities, as the models are able to generalise based on the vast amount of learned patterns and recognise correlations that traditional ML models may not be capable of. These AI-based analysers may enable significant automation of large-scale software system audits or even detect hidden, sophisticated attacks, especially if they are able to learn from past attack patterns, such as the characteristics of supply chain malware.

Machine learning-based static code analysis can offer significant advances in the protection of cyber-physical systems, especially when integrated into continuous integration processes, where fast, automated and reliable decision-making is critical. Such approaches allow a dedicated ML-based verification step to identify potentially problematic patterns before code or updates are deployed, potentially enabling early detection of zero-day vulnerabilities. In the automatic analysis of supplier code, this type of method is not only suitable for searching for specific defects, but may also be capable of filtering out suspicious components based on broader behavioural patterns. This significantly reduces the risk of intentional tampering, before the code is deployed in a production environment.

Conclusion

The increasing digitisation, interconnectivity, and convergence of information technology and operational technology environments have significantly expanded the

[41] Wang et al. 2021.
[42] Li 2025.

cybersecurity attack surface of critical infrastructures. Legacy systems, which were often designed without robust security considerations, are now exposed to sophisticated and evolving cyber-physical threats with potentially severe societal, economic and national security implications. Due to the multifaceted nature of cyber-physical systems, traditional security approaches are proving inadequate, creating a growing potential for countermeasures based on cognitive computing. Despite this potential, their practical application within CIs remains limited and lags behind their more mature application in traditional IT domains.

A key obstacle to the advancement of cognitive security solutions in OT environments is the scarcity of representative, real-world datasets possibly due to national security concerns. This limits the generalisability, adaptability and robustness of existing models, especially in detecting zero-day attacks, coping with protocol diversity and the complex operational characteristics of CPS. Although multiple approaches have shown high accuracy, challenges persist in maintaining low false positive rates and extending detection capabilities to encompass a broader range of physical and hybrid attack vectors.

Additionally, while much of the existing cognitive-computing-based security research has been developed in IT environments, these methodologies often rely on open architectures and large-scale data infrastructures that are not typically found in critical infrastructure environments. Nevertheless, with appropriate adaptations, these approaches offer valuable insights and technological foundations that can be transferred to CPS systems, particularly as the boundary between IT and OT continues to blur.

Future research should prioritise bridging the gap between IT and OT cybersecurity by adapting proven IT-based cognitive computing approaches to the distinct requirements of critical infrastructure. This entails developing robust, privacy-preserving datasets, enhancing model interpretability, and creating context-aware systems tailored to the physical and operational constraints of CIs. Promoting interdisciplinary integration between IT and OT domains is critical to implementing secure, resilient and adaptive defence strategies for critical infrastructures.

References

Alkahtani, Hasan – Aldhyani, Theyazn H. H. (2022): Developing Cybersecurity Systems Based on Machine Learning and Deep Learning Algorithms for Protecting Food Security Systems: Industrial Control Systems. *Electronics,* 11(11), 1–25. Online: https://doi.org/10.3390/electronics11111717

Aragonés, Mario Lozano – Llopis, Israel Pérez – Domingo, Manuel Esteve (2023): Threat Hunting Architecture Using a Machine Learning Approach for Critical Infrastructures Protection. *Big Data and Cognitive Computing,* 7(2), 1–26. Online: https://doi.org/10.3390/bdcc7020065

Assenza, Giacomo – Faramondi, Luca – Oliva, Gabriele – Setola, Roberto (2020): Cyber Threats for Operational Technologies. *International Journal of System of Systems Engineering,* 10(2), 128–142. Online: https://doi.org/10.1504/IJSSE.2020.109127

Baezner, Marie – Robin, Patrice (2017): *Hotspot Analysis: Stuxnet.* Zürich: Center for Security Studies, 1–16. Online: https://css.ethz.ch/content/dam/ethz/special-interest/gess/cis/center-for-securities-studies/pdfs/Cyber-Reports-2017-04.pdf

Bakalos, Nikolaos – Voulodimos, Athanasios – Doulamis, Nikolaos – Doulamis, Anastasios – Ostfeld, Avi – Salomons, Elad – Caubet, Juan – Jimenez, Victor – Li, Pau (2019): Protecting Water Infrastructure from Cyber and Physical Threats: Using Multimodal Data Fusion and Adaptive Deep Learning to Monitor Critical Systems. *IEEE Signal Processing Magazine,* 36(2), 36–48. Online: https://doi.org/10.1109/MSP.2018.2885359

Bondada, Mahesh Babu – Bhanu, S. Mary Saira (2014): Analyzing User Behavior Using Keystroke Dynamics to Protect Cloud from Malicious Insiders. *2014 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM),* 1–8. Online: https://doi.org/10.1109/CCEM.2014.7015481

Booth, Adrian – Dhingra, Aman – Heiligtag, Sven – Nayfeh, Mahir – Wallance, Daniel (2019): *Critical Infrastructure Companies and the Global Cybersecurity Threat.* New York: McKinsey & Company. Online: www.mckinsey.com/~/media/mckinsey/business%20functions/risk/our%20insights/critical%20infrastructure%20companies%20and%20the%20global%20cybersecurity%20threat/critical-infrastructure-companies-and-the-global-cybersecurity-threat-vf.pdf?shouldIndex=false

Budžys, Arnoldas – Kurasova, Olga – Medvedev, Viktor (2024): Deep Learning-based Authentication for Insider Threat Detection in Critical Infrastructure. *Artificial Intelligence Review,* 57(10), 1–35. Online: https://doi.org/10.1007/s10462-024-10893-1

Daricili, A. Burak – Çelik, Soner (2022): National Security 2.0: The Cyber Security of Critical Infrastructure. *Perceptions: Journal of International Affairs,* 26(2), 259–276. Online: https://dergipark.org.tr/en/download/article-file/2181981

Ghafir, Ibrahim – Saleem, Jibran – Hammoudeh, Mohammad – Faour, Hanan – Prenosil, Vaclav – Jaf, Sardar – Jabbar, Sohail – Baker, Thar (2018): Security Threats to Critical Infrastructure: The Human Factor. *The Journal of Supercomputing,* 74(10), 4986–5002. Online: https://doi.org/10.1007/s11227-018-2337-2

Haig, Zsolt – Kovács, László (2012): Kritikus infrastruktúrák és a kritikus információs infrastruktúrák elleni fenygetettségek, támadások. In Ványa, László (ed.): *Kritikus infrastruktúrák és kritikus információs infrastruktúrák.* Budapest: Nemzeti Közszolgálati Egyetem, 92–167.

Hurst, William – Shone, Nathan (2024): Critical Infrastructure Security: Cyber-threats, Legacy Systems and Weakening Segmentation. In Tekinerdogan, Bedir – Akşit, Mehmet – Catal, Cagatay – Hurst, William – Alskaif, Tarek (eds.): *Management and Engineering of Critical Infrastructures.* Cambridge, Mass.: Academic Press, 265–286. Online: https://doi.org/10.1016/B978-0-323-99330-2.00010-6

ISACA (2024): *State of Cybersecurity 2024. Global Update on Workforce Efforts, Resources, and Cyberoperations.* Schaumburg: ISACA. Online: www.isaca.org/resources/reports/state-of-cybersecurity-2024

ISC2 (2024): *Global Cybersecurity Workforce Prepares for an AI-Driven World.* Alexandria: ISC2. Online: https://edge.sitecorecloud.io/internationf173-xmc4e73-prod-

bc0f-9660/media/Project/ISC2/Main/Media/documents/research/2024-ISC2-WFS.pdf

Jermalavičius, Tomas – Rõigas, Henry – Sukhodolia, Oleksandr – Teperik, Dmitri (2025): *The Staying Power of Ukrainian Lights. Lessons of Wartime Resilience of the Electricity Sector.* Tallinn: International Centre for Defence and Security. Online: https://icds.ee/wp-content/uploads/dlm_uploads/2025/05/ICDS_Report_The_Staying_Power_of_Ukrainian_Lights_Jermalavicius_Roigas_Sukhodolia_Teperik_May_2025.pdf

Leszczyna, Rafal – Egozcue, Elyoenai (2013): ENISA Study: Challenges in Securing Industrial Control Systems. In Laing, Christopher – Badii, Atta – Vickers, Paul (eds.): *Securing Critical Infrastructures and Critical Control Systems. Approaches for Threat Protection.* Hershey: IGI Global, 105–143. Online: https://doi.org/10.4018/978-1-4666-2659-1.ch005

Li, Ziyang – Dutta, Saikat – Naik, Mayur (2025): LLM-Assisted Static Analysis for Detecting Security Vulnerabilities. *ArXiv,* 1–24. Online: https://doi.org/10.48550/arXiv.2405.17238

Lis, Piotr – Mendel, Jacob (2019): Cyberattacks on Critical Infrastructure: An Economic Perspective. *Economics and Business Review,* 5(2), 24–47. Online: https://doi.org/10.18559/ebr.2019.2.2

Loukas, George (2015): *Cyber-Physical Attacks. A Growing Invisible Threat.* Newton: Butterworth-Heinemann. Online: https://doi.org/10.1016/B978-0-12-801290-1.00011-4

Markevych, Michal – Dawson, Maurice (2023): A Review of Enhancing Intrusion Detection Systems for Cybersecurity Using Artificial Intelligence (AI). *International Conference, The Knowledge-Based Organization,* 29(3), 30–37. Online: https://doi.org/10.2478/kbo-2023-0072

Mesadieu, Frantzy – Torre, Damiano – Chennamaneni, Anitha (2024): Leveraging Deep Reinforcement Learning Technique for Intrusion Detection in SCADA Infrastructure. *IEEE Access,* 12, 63381–63399. Online: https://doi.org/10.1109/ACCESS.2024.3390722

MITRE Corporation (2024a): *Industroyer.* Online: https://attack.mitre.org/software/S0604/

MITRE Corporation (2024b): *Triton.* Online: https://attack.mitre.org/software/S1009/

MITRE Corporation (2025): *Stuxnet.* Online: https://attack.mitre.org/software/S0603/

Osei-Kyei, Robert – Tam, Vivian – Ma, Mingxue – Mashiri, Fidelis (2021): Critical Review of the Threats Affecting the Building of Critical Infrastructure Resilience. *International Journal of Disaster Risk Reduction,* 60, 1–11. Online: https://doi.org/10.1016/j.ijdrr.2021.102316

Palleti, Venkata Reddy – Adepu, Sridhar – Mishra, Vishrut Kumar – Mathur, Aditya (2021): Cascading Effects of Cyber-attacks on Interconnected Critical Infrastructure. *Cybersecurity,* 4, 1–19. Online: https://doi.org/10.1186/s42400-021-00071-z

Pinto, Andrea – Herrera, Luis-Carlos – Donoso, Yezid – Gutierrez, Jairo A. (2023): Survey on Intrusion Detection Systems Based on Machine Learning Techniques for the Protection of Critical Infrastructure. *Sensors,* 23(5), 1–18. Online: https://doi.org/10.3390/s23052415

Pinto, Andrea – Herrera, Luis-Carlos – Donoso, Yezid – Gutierrez, Jairo A. (2024): Enhancing Critical Infrastructure Security: Unsupervised Learning Approaches for Anomaly Detection. *International Journal of Computational Intelligence Systems,* 17, 1–18. Online: https://doi.org/10.1007/s44196-024-00644-z

Ragsdale, Jarrod – Boppana, Rajendra V. (2023): On Designing Low-Risk Honeypots Using Generative Pre-Trained Transformer Models with Curated Inputs. *IEEE Access,* 11, 117528–117545. Online: https://doi.org/10.1109/ACCESS.2023.3326104

Riggs, Hugo – Tufail, Shahid – Parvez, Imtiaz – Tariq, Mohd – Khan, Mohammed Aquib – Amir, Asham – Vuda, Kedari Vineetha – Sarwat, Arif I. (2023): Impact, Vulnerabilities, and Mitigation Strategies for Cyber-Secure Critical Infrastructure. *Sensors,* 23(8), 1–26. Online: https://doi.org/10.3390/s23084060

Rouzbahani, Hossein Mohammadi – Karimipour, Hadis – Rahimnejad, Abolfazl – Dehghantanha, Ali – Srivastava, Gautam (2020): Anomaly Detection in Cyber-Physical Systems Using Machine Learning. In Choo, Kim-Kwang Raymond – Dehghantanha, Ali (eds.): *Handbook of Big Data Privacy.* Cham: Springer, 219–235. Online: https://doi.org/10.1007/978-3-030-38557-6_10

Sampedro, Carlos – Martinez, Carol – Chauhan, Aneesh – Campoy, Pascual (2014): A Supervised Approach to Electric Tower Detection and Classification for Power Line Inspection. *2014 International Joint Conference on Neural Networks (IJCNN),* 1970–1977. Online: https://doi.org/10.1109/IJCNN.2014.6889836

Shan, Ali – Myeong, Seunghwan (2024): Proactive Threat Hunting in Critical Infrastructure Protection through Hybrid Machine Learning Algorithm Application. *Sensors,* 24(15), 1–24. Online: https://doi.org/10.3390/s24154888

Smeriga, Juraj – Jirsik, Tomas (2019): Behavior-Aware Network Segmentation Using IP Flows. *14th International Conference on Availability, Reliability and Security (ARES 2019),* 1–9. Online: https://doi.org/10.1145/3339252.3339265

Stattelmann, Stefan – Biallas, Sebastian – Schlich, Bastian – Kowalewski, Stefan (2014): Applying Static Code Analysis on Industrial Controller Code. *2014 IEEE Emerging Technology and Factory Automation (ETFA),* 1–4. Online: https://doi.org/10.1109/ETFA.2014.7005254

Tabansky, Lior (2011): Critical Infrastructure Protection against Cyber Threats. *Military and Strategic Affairs,* 3(2), 61–78. Online: www.inss.org.il/wp-content/uploads/sites/2/systemfiles/(FILE)1326273687.pdf

Vávra, Jan – Hromada, Martin – Lukáš, Luděk – Dworzecki, Jacek (2021): Adaptive Anomaly Detection System Based on Machine Learning Algorithms in an Industrial Control Environment. *International Journal of Critical Infrastructure Protection,* 34, 1–11. Online: https://doi.org/10.1016/j.ijcip.2021.100446

Wang, Jingjing – Huang, Minhuan – Nie, Yuanping – Li, Jin (2021): Static Analysis of Source Code Vulnerability Using Machine Learning Techniques: A Survey. *2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD),* 76–86. Online: https://doi.org/10.1109/ICAIBD51990.2021.9459075

Wang, Weiping – Yang, Saini – Hu, Fuyu – Stanley, H. Eugene – He, Shuai – Shi, Mimi (2018): An Approach for Cascading Effects within Critical Infrastructure Systems. *Physica A: Statistical Mechanics and its Applications,* 510, 164–177. Online: https://doi.org/10.1016/j.physa.2018.06.129

Wang, Xuebin – Tan, Qingfeng – Shi, Jinqiao – Su, Shen – Wang, Meiqi (2018): Insider Threat Detection Using Characterizing User Behavior. *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC),* 476–482. Online: https://doi.org/10.1109/DSC.2018.00077

Wilson, Clay (2014): Cyber Threats to Critical Information Infrastructure. In Chen, Thomas M. – Jarvis, Lee – Macdonald, Stuart (eds.): *Cyberterrorism.* New York: Springer, 123–136. Online: https://doi.org/10.1007/978-1-4939-0962-9_7

Zhang, Mu – Chen, Chien-Ying – Kao, Bin-Chou – Qamsane, Yassine – Shao, Yuru – Lin, Yikai – Shi, Elaine – Mohan, Sibin – Barton, Kira – Moyne, James – Mao, Z. Morley (2019): Towards Automated Safety Vetting of PLC Code in Real-World Plants. *2019 IEEE Symposium on Security and Privacy (SP),* 522–538. Online: https://doi.org/10.1109/SP.2019.00034

Zonouz, Saman – Rrushi, Julian – McLaughlin, Stephen (2014): Detecting Industrial Control Malware Using Automated PLC Code Analytics. *IEEE Security & Privacy,* 12(6), 40–47. Online: https://doi.org/10.1109/MSP.2014.113