© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1722890

From Intrusion Detection to Cyberattack Anticipation: A Critical Review of Self-Evolving AI for Pre-Attack Prediction in IoT-Enabled Critical Infrastructure

OPEYEMI PRISCILLA AKINYEMI[1], DAMILARE TIMOTHY OGUNJOBI[2], OBIORAH JOSEPH NWAISAAC[3], RAHMATALLAHI AJIKE ADEROUNMU[4], BENJAMIN OSAZE ENOBAKHARE[5]
[1]*Department of Information and Computer Sciences, University of Luxembourg, Belval, Luxembourg*
[2]*Department of Data Science and Analytics, BI Norwegian Business School, Oslo, Norway*
[3]*Department of Electronics and Computer Engineering, Nnamdi Azikiwe University, Awka, Anambra, Nigeria*
[4]*Department of Zoology, University of Lagos, Lagos, Nigeria*
[5]*Service Support Engineer, Peterbilt Motors (PACCAR Inc), Denton, Texas, USA*

Abstract- ***The growing integration of Internet of Things technologies into critical infrastructure has expanded cyberattack surfaces while increasing the potential consequences of security failures across essential services. Artificial intelligence has substantially improved intrusion and anomaly detection, yet much of this progress remains reactive, identifying malicious activity only after observable evidence of an attack has emerged. This critical review examines whether self-evolving AI can support a transition from intrusion detection to reliable pre-attack anticipation in IoT-enabled critical infrastructure. It distinguishes detection, early warning, forecasting, and anticipation and critically evaluates the role of continual learning, online adaptation, concept drift management, adaptive ensembles, evolutionary approaches, and distributed learning in addressing evolving cyber threats. The review finds that adaptive AI provides important mechanisms for maintaining responsiveness under changing threat and network conditions, but adaptation alone does not constitute predictive intelligence. Evidence for genuine anticipation remains limited by reliance on retrospective benchmark datasets, uncertain pre-attack signals, weak temporal validation, restricted generalizability, and limited evaluation in operational critical infrastructure. Continuous adaptation also introduces challenges involving catastrophic forgetting, adversarial manipulation, false alarms, explainability, computational constraints, and model validation. Progress toward trustworthy anticipation therefore requires prospective temporal evaluation, realistic cross-domain validation, calibrated uncertainty, adversarially robust adaptation, and meaningful human oversight. Future cybersecurity AI should be judged not only by how accurately it detects attacks, but also by whether it can reliably and sufficiently far in advance identify actionable attack risk to enable preventive intervention.***

Keywords: Cyberattack prediction; Self evolving AI; Internet of Things; Critical infrastructure; Intrusion detection; Continual learning; Cybersecurity

I. INTRODUCTION

1.1 IoT-enabled critical infrastructure and emerging cyber risk

Critical infrastructure increasingly depends on Internet of Things technologies for monitoring, automation, communication, and control across energy, healthcare, transportation, water, and industrial systems (Pinto et al., 2023). Although this connectivity improves operational efficiency, it also expands the cyberattack surface by linking heterogeneous devices, networks, cloud platforms, and physical processes. A vulnerability in one connected component may therefore provide an entry point into systems whose disruption can affect essential services and public safety (Feng et al., 2025).

This growing interdependence challenges conventional cybersecurity strategies. Most defensive systems remain reactive, identifying malicious activity through known signatures, anomalous behaviour, or other indicators that become visible during or after an intrusion has begun. Artificial intelligence has improved the speed and accuracy of

such detection, but faster detection does not necessarily constitute prevention. In critical infrastructure, where the interval between compromise and operational disruption may be limited, identifying an attack after its onset may provide insufficient time for intervention. This limitation creates a need to move beyond intrusion detection toward systems capable of identifying credible indicators of cyberattack risk before an attack becomes operational.

1.2 The Detection to Anticipation Gap

Despite rapid advances in artificial intelligence for cybersecurity, much of the existing research remains centred on recognizing attacks once malicious activity is already observable. This creates an important distinction between detection and anticipation. Detection identifies an ongoing or completed intrusion, while early detection reduces the time required to recognize it. Forecasting goes further by estimating the probability of future malicious activity from preceding patterns, whereas pre-attack anticipation requires actionable evidence of elevated attack risk before the attack itself begins.

These capabilities should not be treated as equivalent. A model that detects an intrusion millisecond after onset may offer impressive classification performance without demonstrating any capacity to anticipate it. The critical question is therefore whether AI can move beyond increasingly rapid recognition to identify reliable precursors of attacks with sufficient lead time for preventive action. This detection-to-anticipation gap provides the central analytical problem of this review.

1.3 Self -evolving AI as a proposed solution

Self-evolving AI offers a potential route beyond static cybersecurity models that depend on patterns learned during initial training. In this review, the term refers to AI systems capable of modifying their learned behaviour as new data, threat patterns, and operational conditions emerge. This capability may involve online or continual learning, adaptive learning, and mechanisms for responding to concept drift, where relationships within network data change over time. Such adaptation is particularly relevant to IoT-enabled critical infrastructure because both legitimate system behaviour and cyber threats are dynamic. However, adaptability should not be confused with anticipation. A model that successfully learns from a newly observed attack has demonstrated adaptation, not necessarily an ability to predict that attack. The value of self-evolving AI therefore depends on whether continuous adaptation can support reliable recognition of pre-attack signals rather than merely improve detection of evolving threats.

1.4 Aim and critical questions

This review critically examines whether advances in AI represent a genuine transition from cyberattack detection to pre-attack anticipation in IoT-enabled critical infrastructure. It evaluates the extent to which existing models demonstrate predictive capability before attack onset and whether self-evolving approaches provide meaningful advantages over static detection systems. Particular attention is given to the validity of current experimental evidence and its transferability to operational critical infrastructure. The review further examines the technical, methodological, and governance challenges that may limit reliable anticipation, including model adaptation, generalizability, uncertainty, adversarial manipulation, and human oversight. Through this analysis, the review seeks to distinguish demonstrated predictive capability from claims of anticipation that remain primarily theoretical or detection-based.

II. REVIEW SCOPE AND ANALYTICAL APPROACH

This critical review adopts a purposive and concept-driven approach to examine whether advances in artificial intelligence represent a substantive transition from intrusion detection toward pre-attack cyberattack anticipation in IoT-enabled critical infrastructure. Rather than pursuing exhaustive literature coverage or quantitative evidence aggregation, the review draws on relevant theoretical, methodological, and applied literature to interrogate the conceptual and empirical basis of claims concerning anticipatory cybersecurity.

The literature considered encompasses AI-enabled intrusion and anomaly detection, cyberattack

forecasting and prediction, continual and online learning, concept drift adaptation, adaptive and evolutionary learning, federated and distributed learning, zero-day threat detection, and cybersecurity applications in IoT, industrial control systems, and critical infrastructure. Particular emphasis is placed on studies and perspectives that inform the distinction between detecting malicious activity and identifying actionable attack risk before operational onset.

Evidence was selected purposively according to its relevance to the central analytical problem rather than through exhaustive database screening. Priority was given to literature that contributes to one or more of the following areas: the temporal distinction between detection and prediction; mechanisms through which adaptive AI responds to evolving threats; evidence for pre-attack signals or forecasting capability; methodological limitations of current evaluation practices; and challenges affecting translation to operational critical infrastructure. Seminal contributions were considered where necessary to establish foundational concepts, while recent literature was emphasized to reflect contemporary developments in adaptive and AI-enabled cybersecurity.

The review applies a structured critical lens rather than treating reported model performance as direct evidence of anticipatory capability. Studies and technical approaches are interpreted according to their predictive objective, temporal relationship between model inputs and attack onset, adaptation mechanism, realism of the data and experimental environment, validation strategy, generalizability, treatment of uncertainty, and potential for operational deployment. Particular scrutiny is given to whether information used for prediction would genuinely have been available before attack onset and whether reported outcomes demonstrate forecasting of future risk rather than increasingly rapid recognition of an attack already underway.

The synthesis is therefore interpretive and comparative rather than statistical. Evidence across the literature is integrated to identify where current capabilities are well supported, where findings remain dependent on benchmark or retrospective evaluation, and where claims extend beyond the available empirical evidence. This analytical approach provides the basis for evaluating whether self-evolving AI currently offers credible anticipatory capability and for identifying the methodological and operational requirements necessary to move from adaptive intrusion detection toward trustworthy pre-attack prediction.

III. AI CYBERSECURITY TODAY: THE DOMINANCE OF INTRUSION DETECTION

3.1 Evolution of AI-enabled intrusion detection

Artificial intelligence has progressively shifted intrusion detection from predefined rules toward data-driven recognition of malicious behaviour. Early machine learning approaches improved on signature-based systems by learning patterns associated with normal and abnormal network activity, allowing some previously unseen threats to be identified without an exact known signature. Deep learning extended this capability by extracting complex representations from high-dimensional network and device data, reducing dependence on manually designed features and improving recognition of subtle attack patterns.

More recently, hybrid models have combined complementary learning strategies to improve detection across heterogeneous IoT environments. In parallel, anomaly- and behaviour-based approaches have gained importance because they establish patterns of legitimate activity and identify deviations that may indicate compromise. This is particularly relevant to critical infrastructure, where attack behaviour may evolve and known signatures may provide limited protection against novel threats. However, these advances have largely improved the classification and recognition of malicious activity. They provide much weaker evidence that AI has acquired the temporal reasoning required to anticipate an attack before observable malicious behaviour begins.

3.2 The benchmark performance problem

The apparent progress of AI-based intrusion detection must be interpreted cautiously because

much of the evidence is derived from benchmark datasets rather than operational critical infrastructure. Benchmarks enable reproducible comparison, but they may contain outdated attack patterns, synthetic traffic, simplified network conditions, or class distributions that differ substantially from real IoT environments. Models optimized within these settings can therefore achieve high performance without demonstrating equivalent robustness under changing operational conditions.

Methodological weaknesses further complicate interpretation. Severe class imbalance may favour dominant traffic categories, while inappropriate data partitioning or leakage between training and testing sets can inflate reported performance. Several widely used benchmarks, including the CICIDS2017 corpus, have themselves been shown to contain labelling errors, packet misordering, and other flow-construction artefacts that distort reported detection rates (Engelen et al., 2021). Evaluation also remains heavily centred on accuracy, precision, recall, and related classification metrics. Although useful for measuring detection, these metrics reveal little about prediction horizon, adaptation to changing threats, or performance under previously unseen conditions. More importantly, limited external and prospective validation makes it difficult to determine whether reported gains generalize beyond the datasets on which models were developed. Benchmark superiority should therefore not be treated as evidence of operational readiness or anticipatory capability.

3.3 Why detection is not prediction

A central weakness in the current literature is the tendency to use prediction, detection, forecasting, and anticipation as if they describe the same capability. They do not. Intrusion detection determines whether observed activity is malicious, while early detection reduces the delay between attack onset and recognition. Forecasting instead estimates the likelihood or timing of future malicious activity, and genuine anticipation requires a useful warning before the attack becomes operational.

This distinction changes how model performance should be interpreted. An intrusion detector reporting 99% accuracy may classify attacks exceptionally well while providing no information about whether they could have been identified before onset. Even models described as predictive may rely on features generated during an attack, making the task retrospective classification rather than prospective prediction. Evidence of anticipation therefore requires more than high classification performance. It requires a clearly defined attack onset, information available before that point, a measurable prediction horizon, and validation showing that the warning remains reliable when future attack outcomes are unknown. Without these conditions, claims of pre-attack prediction risk overstating what existing AI systems actually achieve. To clarify this conceptual boundary, Table 1 distinguishes intrusion detection, early warning, forecasting, and pre-attack anticipation according to their temporal position and intended function. This distinction is important because increasing detection speed or accuracy does not, by itself, demonstrate predictive capability.

Table 1. Conceptual distinction between cyberattack detection, early warning, forecasting, and anticipation

Concept	Core function	Temporal position	Key distinction
Intrusion detection	Identifies malicious activity	During or after attack onset	Recognizes an attack already in progress
Early warning	Recognizes emerging malicious activity rapidly	At or shortly after onset	Reduces detection delay but may not predict the attack
Forecasting	Estimates future attack likelihood or timing	Before potential onset	Predicts risk from preceding patterns
Pre-attack anticipation	Identifies actionable evidence of an impending attack	Before attack onset	Seeks sufficient warning for preventive intervention

IV. FROM DETECTION TO PRE-ATTACK ANTICIPATION

4.1 What constitutes a pre-attack signal?

Pre-attack anticipation depends on whether meaningful signals exist before malicious activity reaches the point traditionally recognized as an intrusion. Such signals are unlikely to arise from a single indicator. They may instead emerge from combinations of subtle changes in network behaviour, reconnaissance activity, unusual authentication attempts, vulnerability probing, threat intelligence, and evolving attack paths, consistent with the staged tactics, techniques, and procedures documented in advanced persistent threat research (Alshamrani et al., 2019). Temporal context is equally important because an isolated anomaly may be benign, whereas a sequence of related events may indicate increasing attack intent.

The central difficulty is determining whether these observations are genuine precursors or merely early components of an attack already underway. Reconnaissance and vulnerability scanning, for example, may precede exploitation but can also occur routinely without progressing to compromise. Similarly, authentication anomalies may reflect malicious preparation, legitimate user behaviour, or configuration errors. A useful pre-attack signal must therefore provide information about future risk beyond general abnormality. For critical infrastructure, its value also depends on whether it appears sufficiently early and with sufficient confidence to support preventive action without generating excessive false alarms.

4.2 Predictive approaches

Efforts to move cybersecurity toward prediction increasingly exploit the temporal and relational structure of attack activity. Temporal and sequential models seek to learn how security events develop over time, allowing preceding observations to inform estimates of subsequent malicious behaviour. Behavioural prediction similarly examines changes in user, device, or network activity that may indicate progression toward compromise. Attack graphs provide a complementary perspective by representing possible pathways through which vulnerabilities and system dependencies could be exploited, thereby supporting estimation of plausible future attack states (Pirca & Lallie, 2023).

Threat intelligence can extend these approaches beyond local network observations by incorporating information about emerging vulnerabilities, adversary behaviour, and known attack campaigns. Multimodal models potentially offer a stronger basis for anticipation by combining network traffic, system events, vulnerability information, threat intelligence, and contextual data rather than relying on a single signal source. However, greater model complexity does not automatically produce genuine predictive capability. The critical issue remains whether these approaches learn information that precedes attack onset or simply recognize increasingly subtle evidence that an intrusion has already begun.

4.3 How Strong Is the Evidence for Genuine Anticipation?

Claims of cyberattack prediction require particular scrutiny because the boundary between anticipation and early detection is easily blurred. A model may appear predictive when evaluated against labelled attack sequences while relying on features that become available only after malicious activity has commenced. Under such conditions, strong performance demonstrates recognition of an emerging attack rather than prediction of a future one.

Recent studies illustrate this distinction. Zhai et al. (2026) developed a prediction-based detection and mitigation framework for cyber-physical power systems in which an adaptive Kalman filter forecasts the subsequent evolution of an attack vector. Importantly, however, prediction begins only after the attack has already produced a detectable statistical anomaly in system measurements. The authors explicitly distinguish this capability from pre-attack prediction: the framework forecasts attack progression within the attack timeline, enabling earlier mitigation, but does not anticipate the attack before its first observable manifestation. This provides a useful example of technically meaningful prediction that should nevertheless be classified as

within-attack forecasting rather than genuine pre-attack anticipation.

Other approaches move closer to prospective forecasting by explicitly predicting future attack events or steps. Liu and Guo's (2024) CL-AP² framework, for example, constructs a temporal attack knowledge graph from system logs and uses learned attack trajectories, combining a soft actor-critic algorithm with a transformer model, to predict the attack technique most likely to occur next. Approaches of this kind represent a stronger temporal claim than conventional intrusion classification because their target is a future security event rather than the label of currently observed traffic. Nevertheless, their anticipatory value depends on when the underlying observations become available relative to the true onset of the attack and whether useful warning can be demonstrated outside controlled or retrospectively labelled environments.

This distinction matters because the existence of forecasting models does not imply that all systems described as predictive provide the same temporal capability; the prediction target, information horizon, and temporal position of the input data must be examined explicitly.

Convincing evidence of pre-attack anticipation consequently requires a stricter evidential standard. Model inputs should be restricted to information genuinely available before a clearly defined attack-onset point; the prediction target should concern a future attack event rather than an already evolving intrusion; and performance should report not only discrimination but also the prediction horizon, false-alarm burden, calibration, and operational usefulness of the warning. Evaluation should additionally test whether predictive value persists under temporal change and across environments rather than depending on characteristics of a single benchmark dataset.

The appropriate question is therefore not simply whether an AI system "predicts" attacks, but what it predicts, when the prediction becomes possible, and how much actionable lead time it provides. Until these criteria are demonstrated consistently under prospective and operationally realistic conditions, claims that AI has progressed from intrusion detection to reliable pre-attack anticipation should remain cautious.

V. SELF-EVOLVING AI: THE PROPOSED BRIDGE TO ANTICIPATORY CYBERSECURITY

5.1 What Makes Cybersecurity AI Self-Evolving?

Self-evolving AI is best understood as a functional concept rather than a single algorithmic class. In cybersecurity, it describes systems capable of modifying their learned representations, decision boundaries, or model composition as the threat environment changes, rather than remaining fixed after initial training. Online learning provides one mechanism by updating models as new observations arrive, while continual learning seeks to incorporate new knowledge without losing previously acquired capabilities. Concept drift adaptation addresses changes in the statistical relationships between network behaviour and security outcomes, which is particularly important when legitimate activity and attack strategies evolve.

Other mechanisms extend adaptation beyond individual model updates. Adaptive ensembles can alter the contribution or composition of multiple learners as conditions change, while evolutionary approaches use iterative selection and optimization to refine detection or decision strategies, an approach with an established track record in intrusion detection more broadly (Sen, 2015). Federated and distributed learning may further allow knowledge to be updated across multiple devices or infrastructure nodes without requiring all raw data to be centralized. These mechanisms differ substantially in design and autonomy. What makes them relevant to self-evolving cybersecurity is therefore not the specific algorithm employed, but their capacity to revise learned behaviour in response to new evidence while the security environment continues to change.

5.2 Potential advantages

The principal advantage of self-evolving AI is its potential to reduce dependence on threat patterns fixed during initial training. Continuous adaptation

may allow models to recognize changes in attacker behaviour and legitimate network activity that would progressively weaken static systems. This is particularly relevant to IoT-enabled critical infrastructure, where device populations, communication patterns, software configurations, and vulnerabilities can change during deployment.

Adaptive learning may also strengthen resilience to previously unseen attacks by identifying behavioural deviations without requiring an existing signature for every threat. Distributed and federated approaches could extend this capability across infrastructure nodes, allowing knowledge of emerging patterns to inform multiple devices while limiting the movement of sensitive operational data. More importantly for anticipation, continual observation of temporal changes may enable models to learn sequences that precede particular attack states rather than treating individual events independently.

These advantages, however, remain conditional. Adaptation to new behaviour does not guarantee resilience to genuine zero-day attacks, and knowledge shared across devices may be unreliable when operating contexts differ. Similarly, learning temporal patterns can support forecasting only when stable and informative precursors exist. Continuous adaptation therefore creates opportunities for anticipatory cybersecurity, but does not by itself establish predictive capability.

5.3 Evidence versus promise

The promise of self-evolving AI currently extends beyond what adaptation alone can demonstrate. Existing adaptive systems can update decision rules, respond to concept drift, and incorporate newly observed attack patterns, but these capabilities primarily show that a model can remain effective as its environment changes. They do not establish that the model could have anticipated an attack before malicious activity became observable.

Recent evidence illustrates this distinction. Drift-aware online intrusion-detection systems have demonstrated that models can adjust to changing traffic distributions while maintaining strong detection performance. For example, Makhmudov et al. (2026) combined online Hoeffding Tree classifiers with a leveraging-bagging ensemble and a Page–Hinkley concept-drift test for real-time intrusion detection in Internet of Medical Things networks, reporting an F1-score above 0.99 across binary, medium-scale, and fine-grained attack categories. However, its evaluated objective remains intrusion detection under evolving conditions, rather than forecasting attacks before onset. Such results provide important evidence for adaptive resilience, but not for genuine anticipatory capability.

A similar distinction applies to federated learning. Recent reviews show substantial interest in federated intrusion detection for IoT environments because distributed model training can exploit data across multiple nodes without requiring raw operational data to be centralized. Yet the literature remains predominantly concerned with improving intrusion classification, privacy, communication efficiency, heterogeneous client data, and robustness of distributed learning. A systematic review of federated-learning intrusion-detection systems by Hernandez-Ramos et al. (2025) identified continuing problems in evaluation standardization and emphasized that the field is principally structured around intrusion-detection performance rather than prospective attack anticipation.

Moreover, adaptation introduces vulnerabilities that complicate claims of autonomous predictive improvement. Personalized federated intrusion-detection research by Thein et al. (2024) has demonstrated that poisoning attacks can target the learning process itself, motivating dedicated defensive mechanisms against manipulated updates. Broader IoT federated-learning research likewise identifies heterogeneous local data, privacy protection, model security, and client variability as persistent challenges (Hernandez-Ramos et al., 2025). These findings are especially relevant to self-evolving cybersecurity because continuously incorporating new information increases both the opportunity to learn emerging patterns and the possibility that unreliable or adversarial information will alter future model behaviour.

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1722890

The available evidence therefore supports a more restrained interpretation of self-evolving AI. Online learning demonstrates adaptation to streaming data; concept-drift mechanisms demonstrate maintenance of performance as distributions change; and federated learning demonstrates distributed model updating. None of these functions, by themselves, demonstrates that an impending cyberattack can be identified before attack onset. Genuine anticipation requires an additional evidential step: adaptive models must learn temporally preceding signals and convert them into calibrated estimates of future attack risk within a useful prediction horizon.

Table 2 therefore separates established adaptive functions from their potential contribution to pre-attack prediction.

Table 2. Critical comparison of self-evolving and adaptive AI approaches for cyberattack anticipation

Approach	Adaptation mechanism	Potential contribution to pre-attack anticipation	Key strength	Major limitation	Current evidence limitation
Online learning	Updates the model as new data arrive	May learn emerging precursor patterns as behaviour evolves	Rapid adaptation to changing conditions	Sensitive to noisy, biased, or malicious incoming data	Evidence primarily demonstrates adaptive detection rather than prospective pre-attack prediction
Continual learning	Integrates new knowledge while attempting to retain previous learning	May capture evolving temporal patterns associated with future threats	Supports long-term adaptation	Catastrophic forgetting	Limited evidence that continual adaptation produces reliable warnings before attack onset
Concept-drift adaptation	Detects and responds to changing data relationships	May preserve predictive relevance as network and threat behaviour changes	Maintains model responsiveness under changing conditions	Drift does not necessarily indicate emerging attack risk	Evidence mainly supports maintenance of model performance under drift rather than direct anticipation
Adaptive ensembles	Reweights, replaces, or combines component models as conditions change	May integrate complementary signals associated with developing threats	Combines diverse models and adaptation strategies	Increased computational and operational complexity	Anticipatory benefit remains insufficiently demonstrated under prospective and operational conditions
Evolutionary approaches	Iteratively optimizes models, features, or decision strategies	May discover adaptive patterns relevant to emerging attack states	Flexible optimization and adaptation	Computational cost and uncertain real-time scalability	Evidence is largely experimental, with limited validation of useful pre-attack warning horizons
Federated	Shares model	May exploit	Supports	Heterogeneity	Limited evidence

Approach	Adaptation mechanism	Potential contribution to pre-attack anticipation	Key strength	Major limitation	Current evidence limitation
adaptive learning	knowledge or updates across distributed nodes	distributed precursor signals without centralizing raw data	collaborative learning with greater data locality	and vulnerability to poisoned updates	of prospective anticipation across heterogeneous real-world infrastructure

VI. WHY SELF-EVOLVING AI MAY STILL FAIL IN CRITICAL INFRASTRUCTURE

6.1 Concept drift and catastrophic forgetting

Continuous adaptation creates a fundamental tension between learning new threats and preserving knowledge of existing ones. Concept drift occurs when relationships between network behaviour and security outcomes change over time, potentially reducing the reliability of models trained under earlier conditions (Gama et al., 2014). Adaptive systems can respond by updating their representations, but repeated updating may introduce catastrophic forgetting, whereby learning new patterns degrades performance on previously recognized threats (Parisi et al., 2019; Kirkpatrick et al., 2017). This is particularly problematic in critical infrastructure because older attack strategies do not necessarily disappear when new ones emerge. A self-evolving system must therefore achieve more than rapid adaptation. It must distinguish meaningful environmental change from temporary variation while retaining security knowledge that remains operationally relevant. Without this balance, adaptation could replace one form of model obsolescence with another.

6.2 Adversarial Manipulation and Data Poisoning

The capacity to learn continuously also creates an additional attack surface. If model updates depend on incoming operational data, adversaries may attempt to manipulate those data so that malicious behaviour is gradually interpreted as normal or legitimate activity is misclassified as threatening (He et al., 2023). Data poisoning is particularly concerning because corruption introduced during adaptation may persist within subsequent model decisions (Thein et al., 2024). Distributed learning introduces further risks when compromised devices contribute misleading updates to a shared model. In critical infrastructure, such manipulation could undermine both detection and anticipation while preserving the appearance of normal model operation. Self-evolution therefore requires mechanisms for verifying the integrity of new information and controlling when adaptation is permitted. Otherwise, the ability to learn from changing environments may also become an ability to learn from an attacker's deliberate deception.

6.3 False alarms and uncertainty

Pre-attack prediction introduces a different risk from conventional detection because decisions may be required before malicious intent is certain. In critical infrastructure, a false alarm could prompt unnecessary isolation of devices, interruption of services, or other defensive actions with operational consequences. Conversely, an overly cautious model may fail to identify a developing threat. Predictive performance should therefore not be judged solely by whether an eventual attack is correctly classified. The confidence and uncertainty associated with each forecast, the expected cost of intervention, and the consequences of missed attacks must also be considered. For anticipatory AI to be operationally useful, it must provide sufficiently calibrated risk estimates to support proportionate decisions rather than simply generate additional alerts.

6.4 Explainability and human oversight

Anticipatory systems must provide more than a prediction if operators are expected to act before an attack is confirmed. Complex adaptive models may generate risk estimates without making clear which behaviours, vulnerabilities, or contextual signals produced the warning. This opacity is especially

problematic in critical infrastructure, where preventive actions may affect essential services and require rapid justification. Explainable intrusion-detection research argues that interpretability is central to building operator trust and enabling meaningful oversight of otherwise opaque adaptive models (Neupane et al., 2022). Explainability should therefore help operators assess the evidence supporting a forecast, its uncertainty, and the consequences of intervention. Human oversight remains necessary to determine whether an automated warning warrants action, particularly when evidence is ambiguous. As models evolve after deployment, explanations must also remain meaningful as their decision processes change. Anticipation without sufficient interpretability risks replacing delayed intervention with poorly justified intervention.

6.5 Edge resource and latency constraints

The computational demands of continuous learning may conflict with the resource limitations of IoT environments. Many edge devices operate with restricted processing capacity, memory, energy, and bandwidth, while critical infrastructure may contain heterogeneous hardware with different communication and security capabilities. Benchmarking of continual learning methods on real resource-constrained hardware confirms that memory footprint, latency, and energy consumption can vary sharply across architectures and often narrow the range of deployable options well below what accuracy figures alone would suggest (Li et al., 2025). Models requiring frequent retraining or extensive data exchange may therefore perform well experimentally but prove difficult to sustain in deployment. Cloud-based processing can reduce local computational demands, yet dependence on remote infrastructure may introduce latency, connectivity, and data governance concerns. Anticipatory systems must consequently balance model complexity with the speed and resource efficiency required to produce warnings while preventive action remains possible.

6.6 The validation paradox

Continuous adaptation also complicates the validation on which critical infrastructure security depends. A model validated before deployment may subsequently alter its behaviour as it learns from new data, meaning that its operational state can diverge from the version originally tested. Repeated certification after every meaningful update would be difficult, yet unrestricted adaptation could allow performance or safety to deteriorate without immediate recognition. Monitoring must therefore extend beyond predictive accuracy to include changes in model behaviour, uncertainty, and decision boundaries over time, consistent with broader recognition that production machine learning systems require continuous monitoring of performance, input data, and explanations rather than one-time validation at deployment (Klaise et al., 2020). This produces a central paradox for self-evolving cybersecurity: the adaptability that may preserve effectiveness against changing threats can simultaneously make the system harder to validate, govern, and trust. Reliable anticipation will therefore require controlled evolution rather than unrestricted autonomy.

VII. THE REALITY GAP: FROM BENCHMARK SUCCESS TO CRITICAL INFRASTRUCTURE DEPLOYMENT

The transition from experimental performance to operational critical infrastructure represents a major test for anticipatory AI. Models are commonly developed and evaluated within bounded datasets in which network structure, attack categories, device behaviour, and ground truth are already defined. Operational infrastructure is considerably less controlled. Smart grids and industrial control systems combine cyber and physical processes with strict timing and availability requirements, while healthcare environments contain diverse connected devices alongside sensitive clinical information. Transportation, water, and smart city systems introduce further variation in architecture, communication protocols, device density, and acceptable levels of operational disruption. These differences affect both the signals available to an AI system and the consequences of incorrect predictions. Generalizability is therefore more difficult than demonstrating high performance on an individual IoT dataset. A model may learn relationships that are specific to particular devices, traffic distributions,

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1722890

attack scenarios, or network configurations rather than transferable characteristics of emerging cyber threats — a limitation long recognized in the zero-day detection literature, where models trained on known attack families routinely fail to generalize to genuinely unseen ones (Guo, 2023). This problem becomes more significant for pre-attack prediction because precursor signals may be highly dependent on context. Behaviour that indicates reconnaissance in one environment may represent routine activity in another, while the temporal progression of an attack against an industrial controller may differ substantially from one targeting a connected healthcare device.

Self-evolving AI could partly address this problem by adapting to local conditions after deployment, but adaptation does not eliminate the need for external validation. Indeed, extensive local adaptation may make performance increasingly environment-specific. Evidence of operational readiness should therefore include evaluation across infrastructure domains, changing network conditions, previously unseen attack scenarios, and prospective data rather than repeated optimization on established benchmarks. The relevant standard is not whether a model performs successfully in one controlled environment, but whether its predictive value remains sufficiently reliable when the devices, threats, operational constraints, and consequences of error differ from those represented during development.

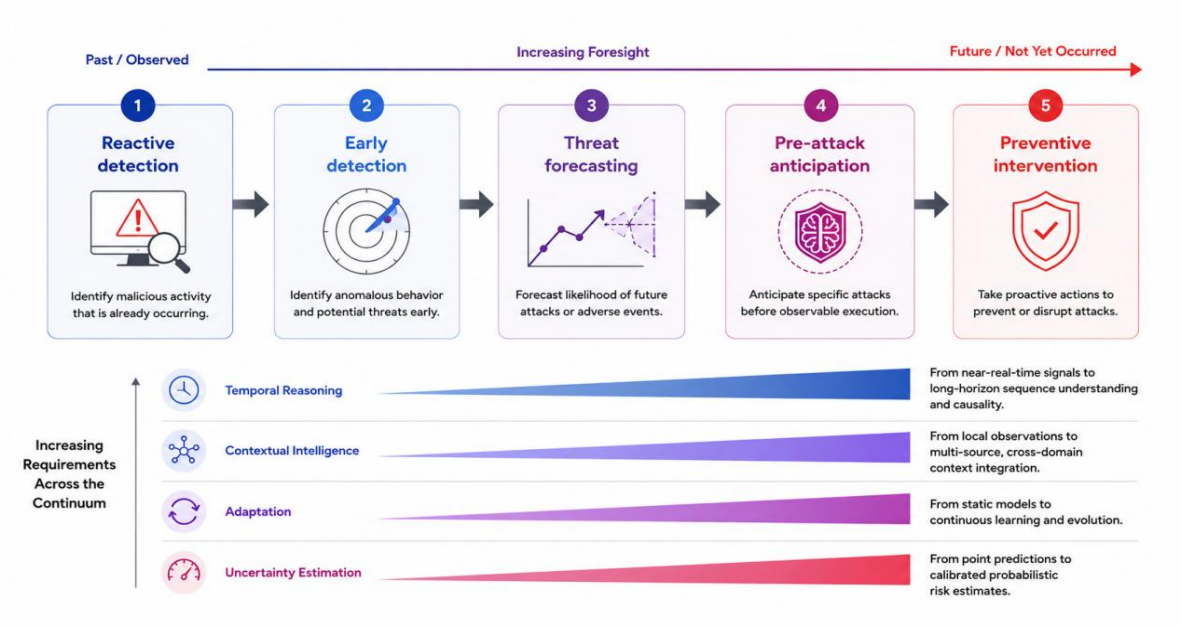

Figure 1. The Detection to Anticipation Continuum in AI-Enabled Critical Infrastructure Cybersecurity. The framework illustrates the progression from reactive intrusion detection to preventive intervention through early detection, threat forecasting, and pre-attack anticipation. Movement toward anticipation requires increasing temporal reasoning, contextual intelligence, adaptive capability, and uncertainty estimation.

VIII. A RESEARCH AGENDA FOR TRUSTWORTHY CYBERATTACK ANTICIPATION

Progress toward trustworthy cyberattack anticipation requires a change in how predictive cybersecurity systems are developed and evaluated. The immediate priority is to move beyond retrospective benchmark classification toward prospective evidence. Datasets should preserve the temporal sequence preceding, during, and following attacks so that models can be

tested using only information that would genuinely have been available at the time of prediction. Longitudinal data are equally important for determining whether performance persists as network behaviour, vulnerabilities, and adversarial strategies change. This requires clearer definitions of attack onset, prediction horizon, and pre-attack information. Without such standards, comparisons between studies will remain difficult, and early detection may continue to be reported as prediction.

Evaluation must also move closer to operational conditions. Models should be tested across different devices, network configurations, attack scenarios, and critical infrastructure domains rather than repeatedly optimized on individual benchmarks. Continual learning benchmarks should specifically assess whether adaptation improves performance under evolving threats without degrading previously acquired knowledge. External and cross-domain validation would further reveal whether models learn transferable indicators of attack risk or merely reproduce characteristics of their development environment. For anticipatory systems, conventional classification metrics should be complemented by measures of prediction horizon, calibration, false alarm burden, robustness under concept drift, and the operational consequences of incorrect forecasts.

Trustworthiness must develop alongside predictive capability. Adaptive models should be evaluated against poisoning and adversarial manipulation because a system that continuously learns also provides opportunities for its future behaviour to be influenced. Predictions should communicate calibrated uncertainty and provide explanations that allow operators to understand the evidence underlying an alert. Human oversight is particularly important where preventive responses could disrupt essential services, echoing broader calls for human agency and oversight as core requirements of trustworthy AI throughout the system lifecycle (Liu et al., 2021; Ottun & Flores, 2025). Future systems should therefore support human judgement rather than treat autonomy as an objective in itself.

A stronger evaluation principle follows from these requirements. Claims of cyberattack anticipation should demonstrate prediction before attack onset, a practically useful prediction horizon, calibrated confidence, adaptation to evolving threats, external validation, and operational feasibility. Meeting these conditions would provide substantially stronger evidence of anticipatory capability than high classification accuracy alone.

IX. CONCLUSION

Artificial intelligence has substantially advanced intrusion detection by improving the recognition of complex, evolving, and previously unseen patterns of malicious activity. Yet this progress should not be mistaken for an equivalent advance in cyberattack anticipation. Much of the current evidence still demonstrates increasingly sophisticated detection rather than reliable prediction before attack onset. The distinction is particularly important in IoT-enabled critical infrastructure, where the practical value of prediction depends on whether sufficient warning is available to prevent or limit disruption.

Self-evolving AI offers a plausible route toward narrowing this gap. Continual learning, concept drift adaptation, distributed learning, and related mechanisms may allow security models to remain responsive as networks and threats change. However, adaptability is not predictive intelligence. Its value for anticipation will depend on whether evolving models can identify genuine precursor signals, estimate future risk with calibrated uncertainty, and maintain reliability under adversarial and operational change.

The transition from detection to anticipation therefore requires more than increasingly complex models or higher benchmark accuracy. Prospective temporal evidence, realistic validation, robustness to manipulation, generalizability, explainability, and controlled human oversight are necessary before anticipatory systems can be trusted in critical infrastructure. The central question for the next generation of cybersecurity AI should consequently shift from how accurately attacks can be classified to how reliably, how far in advance, and with what degree of actionable confidence an emerging attack risk can be identified. That shift represents the

substantive boundary between faster detection and genuine cyberattack anticipation.

Declaration

Conflict of Interest

The authors declare that they have no conflicts of interest related to this work.

Funding

The authors received no specific funding or financial support for the preparation of this manuscript.

Author Contributions

All authors contributed to the conception and design of the review, literature analysis and interpretation, drafting and critical revision of the manuscript, and approved the final version for submission.

Data Availability Statement

No new datasets were generated or analyzed for this review. All information synthesized in this manuscript was obtained from publicly available published literature and cited appropriately.

Ethics Approval

Ethical approval was not required for this study because it involved the analysis of previously published literature and did not involve human participants, animals, or identifiable personal data.

REFERENCES

[1] Alshamrani, A., Myneni, S., Chowdhary, A., & Huang, D. (2019). A survey on advanced persistent threats: Techniques, solutions, challenges, and research opportunities. IEEE Communications Surveys & Tutorials, 21(2), 1851–1877. https://doi.org/10.1109/COMST.2019.2891891

[2] Engelen, G., Rimmer, V., & Joosen, W. (2021). Troubleshooting an intrusion detection dataset: The CICIDS2017 case study. In 2021 IEEE Security and Privacy Workshops (SPW) (pp. 7–12). IEEE. https://doi.org/10.1109/SPW53761.2021.00009

[3] Feng, Y., Huang, R., Zhao, W., Yin, P., & Li, Y. (2025). A survey on coordinated attacks against cyber–physical power systems: Attack, detection, and defense methods. Electric Power Systems Research, 241, 111286. https://doi.org/10.1016/j.epsr.2024.111286

[4] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), Article 44. https://doi.org/10.1145/2523813

[5] Guo, Y. (2023). A review of machine learning-based zero-day attack detection: Challenges and future directions. Computer Communications, 198, 175–185. https://doi.org/10.1016/j.comcom.2022.11.001

[6] He, K., Kim, D. D., & Asghar, M. R. (2023). Adversarial machine learning for network intrusion detection systems: A comprehensive survey. IEEE Communications Surveys & Tutorials, 25(1), 538–566. https://doi.org/10.1109/COMST.2022.3233793

[7] Hernandez-Ramos, J. L., Karopoulos, G., Chatzoglou, E., Kouliaridis, V., Marmol, E., Gonzalez-Vidal, A., & Kambourakis, G. (2025). Intrusion detection based on federated learning: A systematic review. ACM Computing Surveys, 57(12), Article 309. https://doi.org/10.1145/3731596

[8] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114(13), 3521–3526. https://doi.org/10.1073/pnas.1611835114

[9] Klaise, J., Van Looveren, A., Cox, C., Vacanti, G., & Coca, A. (2020). Monitoring and explainability of models in production. Workshop on Challenges in Deploying and Monitoring Machine Learning Systems, ICML 2020. https://doi.org/10.48550/arXiv.2007.06299

[10] Li, S., Kwon, Y. D., Lee, L.-H., & Hui, P. (2025). MetaCLBench: Meta continual learning benchmark on resource-constrained edge devices. arXiv preprint. https://doi.org/10.48550/arXiv.2504.00174

[11] Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Liu, Y., Jain, A. K., & Tang, J. (2021). Trustworthy AI: A computational perspective. ACM Transactions on Intelligent Systems and Technology, 14(1), 1–59. https://doi.org/10.48550/arXiv.2107.06641

[12] Liu, Y., & Guo, Y. (2024). CL-AP²: A composite learning approach to attack prediction via attack portraying. Journal of Network and Computer Applications, 230, 103963. https://doi.org/10.1016/j.jnca.2024.103963

[13] Makhmudov, F., Juraev, G., Yusupov, O., Nasriddinova, P., & Kilichev, D. (2026). Drift-aware online ensemble learning for real-time cybersecurity in Internet of Medical Things networks. Machine Learning and Knowledge Extraction, 8(3), Article 67. https://doi.org/10.3390/make8030067

[14] Neupane, S., Ables, J., Anderson, W., Mittal, S., Rahimi, S., Banicescu, I., & Seale, M. (2022). Explainable Intrusion Detection Systems (X-IDS): A survey of current methods, challenges, and opportunities. IEEE Access, 10, 112392–112415. https://doi.org/10.1109/ACCESS.2022.3216617

[15] Ottun, A. R. O., & Flores, H. (2025). Trustworthy AI in practice: A comprehensive review of human oversight and human-in-the-loop approaches. TechRxiv. https://doi.org/10.36227/techrxiv.176118749.93102582/v1

[16] Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. Neural Networks, 113, 54–71. https://doi.org/10.1016/j.neunet.2019.01.012

[17] Pinto, A., Herrera, L.-C., Donoso, Y., & Gutiérrez, J. A. (2023). Survey on intrusion detection systems based on machine learning techniques for the protection of critical infrastructure. Sensors, 23(5), 2415. https://doi.org/10.3390/s23052415

[18] Pirca, A. M., & Lallie, H. S. (2023). An empirical evaluation of the effectiveness of attack graphs and MITRE ATT&CK matrices in aiding cyber attack perception amongst decision-makers. Computers & Security, 130, 103254. https://doi.org/10.1016/j.cose.2023.103254

[19] Sen, S. (2015). A survey of intrusion detection systems using evolutionary computation. In Bio-Inspired Computation in Telecommunications (pp. 73–94). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-801538-4.00004-5

[20] Thein, T. T., Shiraishi, Y., & Morii, M. (2024). Personalized federated learning-based intrusion detection system: Poisoning attack and defense. Future Generation Computer Systems, 153, 182–192. https://doi.org/10.1016/j.future.2023.11.028

[21] Zhai, P., Zhang, M., & Wang, X. (2026). Prediction-based attack detection and mitigation mechanism in power systems. Scientific Reports, 16, Article 13252. https://doi.org/10.1038/s41598-026-44076-5