

# Autonomous Defense in Cyber + AI

A Strategic Overview of the Next Generation of Cybersecurity

**Report Date:** February 22, 2026

**Publication:** CyberForge AI Internal Report

# CyberForge AI

## Autonomous Defense in Cyber + AI

*Rights owned by Anudeep, Y. N. Narasimha Rao, K. Veera Venkat*

---

## Executive Summary

---

The cybersecurity landscape is undergoing a paradigm shift, moving from a reactive posture to a proactive, autonomous model. As of early 2026, the proliferation of AI-driven, machine-speed cyberattacks has rendered traditional, human-centric defense mechanisms increasingly inadequate. This report details the rise of Autonomous Cyber Defense (ACD), a new frontier where AI systems not only detect threats but also independently respond, neutralize, and recover from them with minimal human intervention.

The primary drivers for this evolution are the sheer speed, scale, and sophistication of modern attacks, coupled with the chronic shortage of skilled cybersecurity professionals and the pervasive issue of "alert fatigue" within Security Operations Centers (SOCs). ACD leverages a suite of advanced AI technologies—most notably Reinforcement Learning (RL), Machine Learning (ML), and Generative AI—to create intelligent agents capable of defending complex network environments dynamically.

## Key Findings:

---

- **Technological Core:** Reinforcement Learning (RL) has emerged as the leading approach for training autonomous defense agents, utilizing simulated "cyber gyms" to develop and hone their decision-making capabilities. Architectural innovations like entity-based RL and Transformer models are overcoming previous limitations, enabling agents to generalize their skills across diverse and changing network topologies.
- **Operational Impact:** ACD is revolutionizing SOCs by automating incident response, case management, and threat hunting. Case studies, such as DXC Technology's implementation, demonstrate dramatic reductions in analyst workload and response times. The concept of "self-healing networks," particularly in critical sectors like healthcare, is moving from theory to practice, offering unprecedented resilience.
- **Challenges and Governance:** Despite rapid progress, significant hurdles remain. These include technical challenges of scalability and transferability to real-world systems, the "black box" nature of some AI models, and the profound ethical and legal questions surrounding accountability. The "Human-in-the-Loop" (HITL) model remains a non-negotiable component, ensuring human oversight and control over critical decisions.
- **The Dual-Use Dilemma:** The same technologies that empower autonomous defense

## 1. The Evolving Threat Landscape: A Race Against the Machine

---

The fundamental challenge of modern cybersecurity is a mismatch of speed. While defenders have historically operated at human speed, adversaries are now leveraging automation and AI to launch attacks that unfold in milliseconds. This asymmetry has rendered traditional defense strategies obsolete and created an urgent need for a new defensive paradigm.

### 1.1 The Inadequacy of Traditional Defenses

---

For decades, cybersecurity relied on signature-based detection. This approach, used by early antivirus and intrusion detection systems, maintains a database of known malware "signatures." While effective against previously identified threats, it is inherently reactive and completely blind to novel or "zero-day" attacks. As attackers began using polymorphic and metamorphic code to alter their malware's signature with each infection, the efficacy of this model collapsed.

Heuristic and anomaly detection systems represented an improvement, establishing a baseline of "normal" network behavior and flagging deviations. However, these systems are often plagued by high false-positive rates, contributing to a critical problem in modern security operations.

### 1.2 The Rise of AI-Powered Attacks

---

The same AI technologies that promise to revolutionize defense are being weaponized by adversaries. The cybercrime landscape of 2025-2026 is characterized by:

- **AI-Generated Phishing:** Generative AI and Large Language Models (LLMs) are used to craft highly personalized and contextually aware phishing emails, devoid of the grammatical errors that once betrayed them. These AI systems can even engage in real-time, convincing conversations with targets to build trust.
  - **Autonomous Malware:** AI-driven malware can adapt its behavior in real-time to evade detection. It can analyze a network's defenses, identify vulnerabilities, and alter its tactics to remain hidden, a process far too fast for human analysts to counter manually.
  - **Automated Vulnerability Discovery:** Attackers use AI to scan for and discover new
-

## 2. Defining Autonomous Cyber Defense (ACD)

---

While there is no single, universally agreed-upon definition, Autonomous Cyber Defense (ACD) represents a conceptual leap beyond simple automation. It is a paradigm where intelligent agents are empowered to execute the full spectrum of cybersecurity tasks—from detection and analysis to response and recovery—at machine speed and without direct human tasking for each action.

### 2.1 From Automation to Autonomy

---

It is crucial to distinguish between automation and autonomy.

- **Automation**, as seen in Security Orchestration, Automation, and Response (SOAR) platforms, involves executing pre-defined, rule-based playbooks. For example, "If alert X is triggered, then block IP address Y." This is rigid and effective only for known, predictable scenarios.
- **Autonomy**, in contrast, implies decision-making. An autonomous agent uses AI to perceive its environment, understand context, and choose the optimal course of action from a range of possibilities to achieve a goal. It can adapt to novel situations not explicitly covered by a playbook.

*"A new era is emerging: Autonomous Cyber Defense. A paradigm where AI does not just detect threats, it responds, neutralizes, and recovers." - LinkedIn Pulse*

### 2.2 The Full Defense Lifecycle

---

ACD encompasses the entire NIST Cybersecurity Framework, with a particular focus on accelerating post-detection functions. While AI has long been used for `Identify` and `Protect`, true ACD excels in:

1. **Detect:** Using advanced behavioral analytics and ML to identify anomalies and malicious patterns that evade signature-based tools.
  2. **Respond:** This is the core of ACD. Upon detection, an agent can autonomously execute
-

## 3. Core Technologies Powering Autonomy

Autonomous Cyber Defense is not a single technology but an integration of several advanced AI disciplines. Each technique plays a distinct role in creating systems that can perceive, reason, act, and adapt within a complex digital environment.

*Figure 1: ACD research focuses on the 'Respond' and 'Recover' functions of the NIST Cybersecurity Framework, which are historically the most manual and time-consuming. (Source: Adapted from Dstl)*

### 3.1 Reinforcement Learning (RL): The Decision-Making Engine

RL is the leading AI approach for creating the autonomous agents at the heart of ACD. Unlike other forms of machine learning that learn from static data, RL agents learn through trial and error by interacting with an environment.

- **How it Works:** An agent takes an `action` in an `environment` (e.g., a simulated network). It then receives an `observation` (the new state of the network) and a `reward` (a score indicating if the action was beneficial or detrimental to the network's security). The agent's goal is to learn a `policy`—a strategy for choosing actions—that maximizes its cumulative reward over time.
- **From Games to Cyber:** RL gained prominence by mastering complex games like Go and Atari. Researchers are now applying the same principles in "cyber gyms"—simulated network environments like `CybORG` and `Yawning Titan`—where defensive agents can be trained to combat simulated attackers.

### 3.2 Machine Learning (ML): The Pattern Recognition Layer

ML models serve as the sensory system for autonomous agents, analyzing vast datasets to identify patterns indicative of a threat.

- **Supervised Learning:** Trained on labeled data (e.g., "this is malware," "this is a phishing email"), this is used to detect known threats with high accuracy.

## 4. Architectural Innovations for Generalization

---

A primary technical barrier to deploying autonomous agents in the real world has been the problem of `generalization`. An agent trained in one specific, static network simulation often fails when deployed in a real, dynamic enterprise network where devices constantly join and leave. Recent architectural breakthroughs are addressing this challenge head-on.

### 4.1 The Limitation of Fixed-Input Models

---

Standard deep learning models, such as Multi-Layer Perceptrons (MLPs), expect a fixed-size input. In cybersecurity, this meant an agent's "observation" of the network had to be a vector of a constant length. This approach is brittle because:

- It cannot handle networks of different sizes than the one it was trained on.
- The meaning of an input feature is tied to its position in the vector, making it difficult to represent the complex, graph-like relationships between nodes in a network.

An agent trained on a 20-node network simply could not be deployed on a 50-node network without complete retraining.

### 4.2 Entity-Based Reinforcement Learning

---

To overcome this, researchers are reframing the problem using an `entity-based` approach. Instead of a single flat vector, the environment is decomposed into a collection of discrete entities (e.g., nodes, subnets, users).

*"In the context of a network environment, each node of the network can be treated as an entity, with a defending agent's observation space permitted to vary between environment instances according to how many nodes are visible." - Entity-based Reinforcement Learning for Autonomous Cyber Defence, 2024*

This framework allows the agent's policy to be parameterized using architectures designed for compositional generalization, which are invariant to the number and order of entities.

---

## 5. Real-World Applications & Case Studies

---

While fully autonomous systems are still emerging, AI-enhanced automation is already delivering transformative value in security operations. These applications serve as a bridge to a future of greater autonomy, demonstrating tangible benefits in efficiency, accuracy, and resilience.

### 5.1 The AI-Driven Security Operations Center (SOC)

---

AI is augmenting and automating tasks across the SOC, empowering understaffed teams to do more, faster. Key use cases include:

- **Automated Triage:** AI filters the flood of alerts, automatically dismissing false positives and prioritizing genuine threats based on risk scoring, allowing analysts to focus on what matters.
- **Enhanced Case Management:** AI assists junior analysts by suggesting next steps based on historical data and established frameworks like MITRE ATT&CK, effectively upskilling Tier 1 analysts.
- **Streamlined Reporting:** AI generates comprehensive incident summaries and shift-change reports, ensuring seamless transitions and clear communication with stakeholders.
- **Phishing Analysis:** AI inspects suspicious emails, analyzes headers and content, and explains why a message is likely malicious, accelerating response to one of the most common attack vectors.

### 5.2 Case Study: DXC Technology's SOC Revolution

---

DXC Technology, a global IT services leader, manages a massive cybersecurity infrastructure. Facing overwhelming alert volumes, the company implemented an AI-driven SOC strategy to modernize its defenses.

- **Challenge:** Thousands of daily alerts, high rates of false positives, and significant analyst fatigue, leading to slow response times (MTTR).
  - **Solution:** Deployed an agentic AI platform (from partner 7AI) to automate alert
-

## 6. Strategic Initiatives and Research Programs

---

The development of Autonomous Cyber Defense is being driven by a concerted effort from government agencies, academia, and the private sector. These programs aim to accelerate research, foster innovation, and build the foundational technologies for the next generation of cybersecurity.

### 6.1 DARPA: Pioneering Autonomy

---

The U.S. Defense Advanced Research Projects Agency (DARPA) has been a central force in pushing the boundaries of autonomous systems for national security.

- **Cyber Grand Challenge (CGC):** Held in 2016, this landmark competition was the world's first all-machine hacking tournament. It challenged teams to create Cyber Reasoning Systems (CRS) that could autonomously find, patch, and exploit software vulnerabilities in real time. CGC proved that automated, machine-speed defense was possible.
- **Active Cyber Defense (ACD):** This program sought to give defenders a "home field" advantage by developing capabilities to discover, analyze, and mitigate threats in real time. It focused on proactive defense within DoD-controlled cyberspace.
- **Assured Autonomy:** This program focuses on ensuring that learning-enabled, cyber-physical systems (like autonomous vehicles or weapons) are safe, reliable, and trustworthy. It addresses the core challenge of verifying the behavior of complex AI systems.

### 6.2 The CAGE Competition and International Collaboration

---

Building on these foundations, international collaborations are nurturing the fledgling field.

- **Cyber Autonomy Gym for Experimentation (CAGE):** An initiative by The Technical Cooperation Program (TTCP), CAGE provides a standardized framework and simulation environment for developing and testing RL-based defensive agents. Competitions hosted within CAGE help benchmark progress and attract new talent.
- **ARCD Programme (UK):** The UK's Autonomous Resilient Cyber Defence (ARCD) programme, funded by the Defence Science and Technology Laboratory (Dstl), aims to

## 7. The Human-in-the-Loop (HITL) Imperative

---

As autonomous systems become more capable, their relationship with human operators becomes the most critical factor for safe and effective deployment. The goal of autonomy is not to replace human expertise but to augment it, freeing analysts from repetitive, high-volume tasks to focus on strategic decision-making, complex threat hunting, and contextual judgment. The Human-in-the-Loop (HITL) model is non-negotiable, especially in safety-critical systems.

### 7.1 Why HITL is Essential

---

Fully automated systems, while fast, are often rigid and context-blind. They can make high-impact mistakes when faced with ambiguous or novel situations. HITL systems intentionally build human input into the decision-making process to:

- **Provide Context:** Humans bring business context, institutional knowledge, and an understanding of acceptable risk that machines lack. An automated system might block a server to contain a threat, but a human knows if that server runs a life-sustaining hospital application.
- **Reduce False Positives:** Analysts can quickly dismiss benign alerts that an AI flags as anomalous, preventing unnecessary disruption and refining the AI's future performance.
- **Handle Ambiguity:** When an AI agent encounters a situation with high uncertainty, it can escalate to a human for a final judgment call.
- **Ensure Accountability:** Keeping a human involved maintains a clear chain of responsibility for critical actions.

### 7.2 Models of Human-Machine Interaction

---

The "loop" can involve different levels of interaction, forming a spectrum of autonomy:

1. **Human-in-the-Loop (HITL):** The human is directly involved in the decision-making process. The AI may propose an action (e.g., "Isolate this endpoint?"), but it requires explicit human approval before execution. This is common for high-stakes responses.
  2. **Human-on-the-Loop:** The system acts autonomously, but a human supervisor monitors
-

## 8. Challenges and Risks on the Path to Autonomy

---

The path from lab to operational autonomous defense is fraught with significant technical, ethical, and strategic challenges. While the potential benefits are immense, these risks must be proactively managed to ensure the technology is a net positive for society.

### 8.1 Technical Hurdles

---

- **Scalability and Fidelity:** Existing cyber gyms, while useful, are rudimentary compared to the complexity of real-world enterprise networks. Building and maintaining high-fidelity simulations that can train agents for these environments is computationally expensive and requires immense resources.
- **Transferability:** Despite progress with GNNs and Transformers, ensuring that an agent trained in a simulation can perform reliably in a live, constantly changing network (the "sim-to-real" gap) remains a major research problem.
- **Data Scarcity:** Training effective AI models requires vast amounts of high-quality data, including data on real-world attacks and network responses. This data is often sensitive and proprietary, making data sharing for research purposes a delicate issue.
- **Security of the AI Itself:** The autonomous defense system becomes a high-value target. Adversaries will seek to compromise it through:
  - **Data Poisoning:** Maliciously altering the training data to teach the AI to misclassify threats or create backdoors.
  - **Model Evasion:** Crafting attacks specifically designed to be "invisible" to the AI's detection patterns.

### 8.2 The Dual-Use Dilemma

---

One of the most significant strategic risks is that research into autonomous defense inadvertently accelerates the development of autonomous offense. An RL agent trained to find and patch vulnerabilities can be repurposed to find and exploit them.

*"Policymakers should invest in research to determine which scenarios and technologies*

## 9. Ethical and Legal Implications

---

Delegating decision-making authority to machines, especially for actions that can have significant real-world consequences, raises profound ethical and legal questions. As autonomous systems move from defending networks to controlling physical systems (e.g., critical infrastructure, autonomous vehicles), these issues become paramount.

### 9.1 The Accountability Gap

---

Perhaps the most pressing legal challenge is determining liability when an autonomous system acts unpredictably and causes harm. Who is responsible?

- **The Operator?** It may be unreasonable to hold a human operator liable for the unpredictable actions of a "black box" AI they cannot fully understand or control.
- **The Developer/Manufacturer?** Proving negligence in the design of a complex system that evolves after deployment is a significant legal hurdle. The system's behavior may be an "emergent property" not directly intended by its creators.
- **The AI Itself?** Under current legal frameworks, an AI system cannot be held criminally liable as it lacks legal personhood and intent.

This "accountability gap" could leave victims without remedy and erode public trust in AI. Some jurisdictions are beginning to address this; for example, California's AB 316, effective January 1, 2026, precludes using an AI's autonomous operation as a defense against liability claims.

### 9.2 Infringement on Human Rights

---

The deployment of autonomous systems, particularly in law enforcement or military contexts, implicates fundamental human rights. A 2025 Human Rights Watch report highlighted several key concerns regarding autonomous weapons, many of which apply to ACD:

- **Right to Life:** An autonomous system may lack the human judgment needed to assess necessity and proportionality before using force, potentially leading to arbitrary deprivation of life.
  - **Human Dignity:** Delegating life-and-death decisions to a machine that cannot
-

## 10. Recommendations and Future Outlook

---

Autonomous Cyber Defense is a long-term ambition, but the foundational work being done today will shape the security of our digital world for decades. To realize the potential benefits while mitigating the risks, a coordinated, multi-stakeholder approach is required. The following recommendations are synthesized from expert analysis by institutions like CSET, The Alan Turing Institute, and the World Economic Forum.

### 10.1 Recommendations for Maturing the Technology

---

1. **Invest in Scaling Up Simulation:** Progress depends on the ability to train agents in realistic environments. Sustained funding is needed to build and maintain large-scale, high-fidelity network simulations ("testing and training ranges") that incorporate complex scenarios and attacker behaviors. These resources should be made accessible to the broader research community.
2. **Coordinate Data Sharing:** To overcome data scarcity, policymakers should establish frameworks for the secure and anonymized sharing of cyber data about network configurations and observed threats. While delicate, such initiatives benefit all organizations by improving the data used to train defensive models.
3. **Continue to Host Competitions:** Competitions like CAGE are invaluable for benchmarking progress, fostering innovation, and developing future talent. Financial incentives can help attract top researchers from academia and industry.
4. **Prioritize Research Areas:** Not all defense scenarios require autonomy. Research should be guided toward areas where speed and scale are the limiting factors and where autonomy provides the most significant advantage over existing methods.

### 10.2 Recommendations for Governance and Policy

---

1. **Develop Frameworks for Trust and Autonomy Levels:** Policy guidance is needed to set targets for capability and trustworthiness, matched to the risk of the decisions an agent is authorized to make. This could be analogous to the levels of autonomy defined for autonomous vehicles, establishing clear rules for when and where different levels of autonomy are permissible.
-