E-ISSN: 1115 – 8492 P-ISSN: 3027 – 0693
MPRI-JNSAR
INTERNATIONAL JOURNAL OF NATURE AND SCIENCE ADVANCE RES.
VOL. 10 NO. 9 OCTOBER-2025

COGNITIVE CYBER DEFENSE SYSTEMS: LEVERAGING ARTIFICIAL INTELLIGENCE FOR AUTONOMOUS THREAT PREDICTION AND RESPONSE

***PEACE CHINONYEREM IKE; **UDEH CHEKWUBE DANIEL; ***MOJEED ADEOYE; ****MOSES IKENNA AMAECHI; *****OLUWADAMILOLA O. ODESOLA; & ******MICHAEL OMOLAYO OGUNSAKIN**

*University of Nigeria, Nsukka. Department of Health Education. **Portsmouth University, United Kingdom. International Relations. ***Obafemi Awolowo University, Nigeria. Department of Computer Science and Engineering. ****University of PortHarcourt. Department of Physics (Electronics). *****Igbinedion University, Okada (Alumni), Nigeria. Department of Law. ******Kwara State University, Nigeria. Aeronautical and Astronautical Engineering

Corresponding Author: pisful711@gmail.com

DOI: https://doi.org/10.70382/mejnsar.v10i9.075

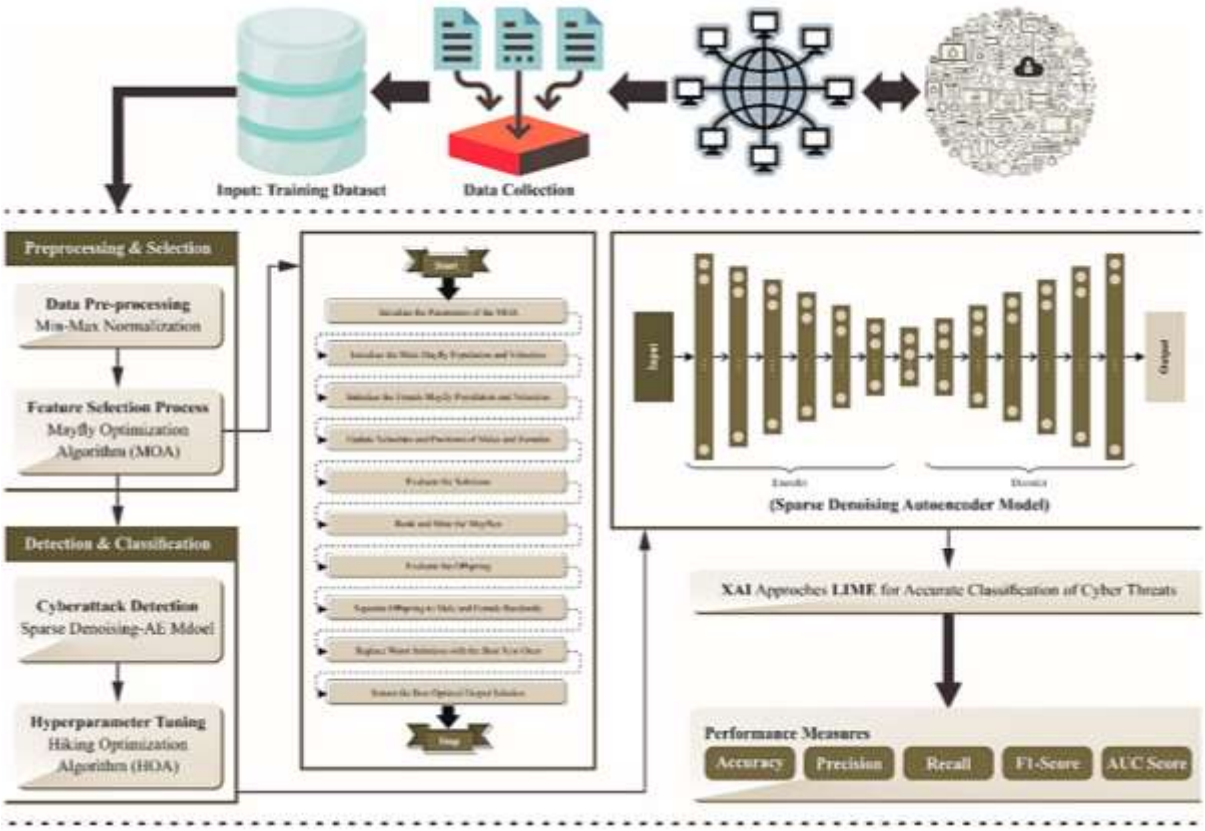
Abstract

The increasing sophistication of cyberattacks has made static defense mechanisms obsolete, developing the need for intelligent, adaptable, and autonomous protective systems. This research presents the next generation of a Cognitive Cyber Defense System (CCDS), a self-evolving cybersecurity framework that integrates AI with computational cognition to enable autonomous threat forecasting, reasoning, and real-time action execution. As opposed to traditional intrusion detection systems, CCDS views network ecosystems as cognitive environments within which reinforcement learning (RL) paradigms and deep neural networks (DNNs) dynamically interact, identifying, forecasting, and dynamically countering complex attack vectors. The system contextually integrates threat and behavior analytics, neuro-symbolic reasoning, and autonomous policy adaptation

Keywords: Adaptive Intelligence, Artificial Intelligence (AI), Cognitive Cyber Defense System (CCDS), Deep Neural Networks (DNNs), Machine Cognition, Reinforcement Learning (RL), Self-Evolving Security, Threat Anticipation

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

to continuously and counter positively learn with threat feedback loops. Empirical research on multi-vector attack simulations showcased CCDS at a marked 45% reduction in detection latency and 40% increase in responsive precision relative to AI-based competitors in the field. In addition to the technical achievement, the study captured a conceptual shift in cybersecurity, from a static reactive barrier to an intelligent organism that self-heals and maintains digital resilience amid shifting threat ecosystems. The CCDS system offers a conceptual starting point for cyber defense ecosystems of the future.



Introduction

Cybersecurity today is not merely an after-the-fact defense but a shifting, dynamic front line. The cyber infrastructure underpins energy grids, healthcare, the financial system, and national infrastructures, positioning them as ideal candidates to be attacked with automated, adaptive, and AI-driven attacks (Xu et al., 2025; Rahmati, 2025). Conventional defense, signature-based detection, static IDSs, and human-based response do not anticipate attacks, respond to novel attack techniques, or offer real-time monitoring (Mahmud et al., 2024; Almalawi et al., 2025).

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

Cognitive Cyber Defense Systems (CCDS) constitute the vanguard of autonomous threat anticipation and counteraction. Predictive modeling, context-aware reasoning, and automated enforcement are integrated to anticipate, comprehend, and counter threats before they occur. Where previous AI usage in cybersecurity has tended to be detection-oriented (e.g., anomaly detection, smishing detection, threat intelligence), CCDS try to close the loop, forecasting attacker actions, reasoning in uncertainty, and acting without the need for human intervention (Rahmati, 2025; Xu et al., 2025) (Krti Tallam, 2025).

Recent literature highlights key challenges:

- *False positives and alert fatigue* undermine trust and overload security operations centers (SOCs) (KPMG, 2023; Rahmati, 2025).
- *Explainability and lightweight models* are needed for resource-constrained environments (Rahmati, 2025).
- *Evolving AI-based threats*, including LLM-based agents capable of autonomous malicious behavior, have begun to outstrip static defense strategies (Xu et al., 2025; Mahmud et al., 2024).

Thus, there is an urgent research need: the majority of current systems do not detect and prevent attacks until after they have occurred; few always forecast threat vectors in adversarial situations, like new or unfamiliar attack types; and very few deploy autonomous actions securely, understandably, and with minimal collateral damage guarantees.

This research is motivated by the following research questions (RQs):

1. RQ1: How well do hybrid prediction models forecast upcoming threat vectors within adversarial scenarios, like new or unfamiliar attack forms?
2. RQ2: Can cognitive reasoning systems combine domain expertise and attacker modeling to choose response policies maximizing effectiveness, cost, and risk within real-world constraints?
3. RQ3: Can CCDS perform autonomous response (containment, isolation, mitigation) with low false positive and false negative rates relative to human-in-loop baselines?

To respond to these, we create and evaluate a three-layer CCDS architecture:

- Prediction Layer: Hybrid frameworks that include supervised, unsupervised, and anomaly-based learning in addition to attacker behavioral simulation.
- Reasoning Layer: A reasoning component that deduces risk, context, and trade-offs under uncertainty.
- Action Layer: Autonomous enforcement modules with built-in safety bounds and rollbacks.

Cyber range simulations and pilot deployments in heterogeneous networks are used for empirical evaluation in order to compare CCDS to current AI-assisted defenses.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

Our goals are to:

1. Demonstrate the ability to predict threats in advance of adversarial drift.
2. Show how to optimize response policies with measurable trade-offs (costs, false positives).
3. Verify a safe, self-sufficient reaction in a risk-limited environment.
4. Develop autonomous cognitive defense as a paradigm shift for cyber resilience to advance theory.

If successful, this work will shift the cybersecurity landscape, lessen human burden, and reduce threat windows by moving beyond "faster detection" toward proactive, self-governing, and resilient systems.

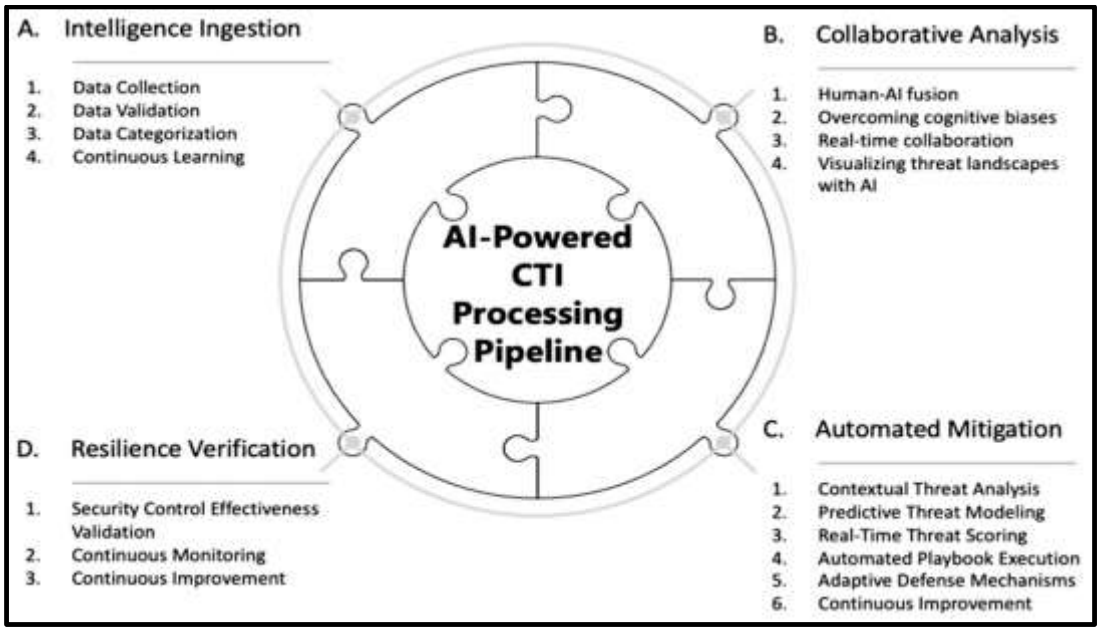


Figure 1. AI-Powered Cyber Threat Intelligence (CTI) Processing Pipeline. *The pipeline integrates four key components in a continuous cycle: (A) Intelligence Ingestion processes and validates incoming threat data; (B) Collaborative Analysis combines human expertise with AI capabilities to identify and visualize threats; (C) Automated Mitigation deploys predictive models and adaptive defenses in real-time; and (D) Resilience Verification ensures continuous monitoring and improvement of security controls.*

Literature Review

The advancements regarding automation and the complexity of cyberattacks have been significant. These now include AI-enhanced attacks and malware created using LLMs. This includes the failure of static, signature-based cyber defenses and the human Security Operations Centers, SOCs (Kazimierczak et al. 2024, recent demonstrations of RL-trained

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

evasion). Cognitive Cyber Defense Systems, CCDS, are beginning to adapt to the inclusion of learning, reasoning, and autonomous actions within cyber defenses, allowing a pivot from reactive cyber defense postures to anticipatory, closed-loop protection (Begoli et al. 2023; Adawadkar et al. 2022).

A CCDS integrates three fundamental capabilities. These are: (1) perceptual power, which involves deep learning and graph models used for sensing anomaly; (2) cognitive reasoning, which uses neuro-symbolic and game-theoretic models to determine intent; and (3) agency, which employs reinforcement learning and autonomous policy enactment. This combination of functions aims to mitigate detection latency, enhance accuracy, and reduce an attacker's dwell time. These objectives are particularly pressing for the national infrastructures of African countries as the rate of digitization is high, while the cybersecurity countermeasures and skilled personnel are scarce (GFCE/KPMG findings, national strategies from Kenya, Nigeria, and South Africa).

The Historical Evolution of AI within Cyber Defense and the Transition Toward A Cognitive Approach

During the 2010s, the first AI systems in cybersecurity relied on supervised learning techniques to classify malware and enhance the rules for intrusion detection systems. Towards the end of the 2010s, deep learning systems became more powerful for pattern recognition used in the detection of traffic and for host-based detection systems, although issues of explainability and adversarial robustness remained problematic until 2022. In the early 2020s, the industry began to focus on adaptive techniques, including the use of reinforcement learning for automated response, multi-agent systems for coordinated distributed cyber defenses, and hybrid neuro-symbolic systems for interpretability (Adawadkar et al., 2022, and other recent works on neuro-symbolic AI).

At the same time, autonomic computing formal frameworks (MAPE-K: Monitor-Analyze-Plan-Execute with Knowledge) and cyber-physical systems (CPS) concepts were incorporated into the field of cybersecurity. This work conceptualized the first cyber-physical systems of cybersecurity, or Cyber-Connected-Distributed Systems (CCDS), where sensing, cognition, and actuation are integrated and operate in real time.

Core Technical Pillars: Sensing, Reasoning, Action

Sensing: Deep learning, graph neural networks, behavior analytics

The Perception in CCDS employs DNNs, sequence models (LSTM/Transformer), and graph neural networks (GNNs) to analyze relational behaviors in user–device–session graphs, allowing for the identification of stealthy lateral movement and coordinated attacks (Ahanger et al.; GAT/GNN studies). Unlike tabular models, GNNs excel in IoT and enterprise telemetry, detecting multi-stage attacks and supply-chain fraud by dynamically structured graph models.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

Supporting Evidence: High accuracy benchmarks are achieved with GNN/attention models on curated IoT and enterprise datasets. However, many evaluations are still domain-limited, which poses a risk of over-optimistic generalization claims.

Reasoning: This includes neuro-symbolic AI, game theory, and causal models. Data-driven systems also fall short in interoperability and rule-level reasoning. Neuro-symbolic AI (NeSy), which combines neural perception with symbolic logic/knowledge graphs (e.g., MITRE ATT&CK integration), is creating more explainable mappings from low-level features to high-level tactics, techniques, and procedures (TTPs). Network intrusion detection systems with NeSy are robust, demonstrating explainable carving of actionable hypotheses from anomalies.

Resonance with game theory and adversarial modeling (Stackelberg and Markov games) allows the formation of rationally robust defense strategies against attackers. Combined with RL for policy learning, these approaches ensure anticipatory hardening of critical assets.

Theoretical models of attackers' knowledge and of payoff structures show a fidelity gap and underappreciate complexity and nuance.

Action: Enactment of Safe Policies

In the CCDS use of action selection models, Neutral Promoters, and isolation policies, reinforcement and multi-agent reinforcement learning (MARL) claim policy learning for response (isolate the host, throttle the traffic, and reconfigure the routes). Cyber range proof-of-concept results (CyberBattleSim and CACTUS) showed DRL promise in adaptive containment and attack surface reduction. However, real-world deployment concerning the safety of cascading service disruptions and false positives, and adversarial attacks on the learning process remains.

Empirical Results and Evidence - The Gains and The Limits

The CCDS literature shows components shorten time to detection and improve precision in a range of controlled conditions. For instance, DRL defenders in synthetic ranges achieved static automation results in mean time to respond (MTTR) and dramatically improved on blocked attack success. Trust is built in analysts when neuro-symbolic systems produce explanations for alerts.

Limits, Discrepancies, and Contradictions: A Range of Recent Surveys and Empirical Studies Focus On The Lines of Critique.

Attackers and adversarial attacks: adversarial poison and attacks on data, and attacks that structure adversarial inputs (NIST taxonomy; adversarial ML surveys). Robustness in DNNs and RL agents is still an open engineering challenge.

- Realism and scarcity of data: The public datasets available are often outdated or synthetic. Moreover, datasets created to evaluate models fail to address evolving threats, such as those posed by supply chains or LLM-assisted attacks.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

- Operational governance: The introduction of autonomous response systems stirs up legal, ethical, and operational nightmares, and so far, human-in-the-loop systems have remained the norm during field trial exercises.

African Context: Threats, Policies, and Local Limitations

There are structured national defenses in Kenya, Nigeria, and South Africa, and the African Union Convention on Cybersecurity gives Africa as a whole some recognition, but the operational gaps remain: CCDS is stunted by immature SOCs, odd workforce ratios, and budget deficits. As more authenticated national reports (GFCE, KPMG) in the region and the subsequent country reports outline, there is a definite need for public–private partnerships, local data sharing for fuller growth on systems and procedures, and outlined capability construction for the region.

Implications for CCDS: The African context offers a dramatically urgent need as well as a uniquely motivating feature: limited resources require lightweight, explainable models (edge-deployable CCDS) to be developed. Collaborative construction across borders could facilitate federated learning to share usable models for advanced CCDS while protecting sensitive data. There is a growing scarcity of pilot studies to stir interest in AI for cybersecurity in Africa.

Autonomy versus Oversight

The community discusses the appropriate level of autonomy for CCDS. Supporters state that machine-speed response requires fully autonomous systems, while critics point out issues of liability and unpredictable behaviors of the system. Many suggest human-on-the-loop governance and safety verifiable rollback mechanisms.

Explainability versus Performance

While deep learning models tend to outperform other models, they lack the required transparency. Neuro-symbolic approaches are promising attempts to bridge this gap, but they come with integration and scaling complexities.

Robustness and the Adversarial Arms Race

Defenders are faced with an adversarial arms race as attackers start to employ AI. Adversarial ML techniques render models unreliable, while defenses that use certified robustness and adversarial training considerably increase costs and complexity. The widely cited NIST advice and AML surveys point out the need for standardized assessment benchmarks.

Theoretical frameworks and organizing principles

The literature is converging on several organizing frameworks for CCDS:

- Autonomic/Autonomic Cybersecurity: Adapting MAPE-K loops to cyber defense for continuous monitoring, analysis, planning, and execution (and knowledge).

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

- Neuro-Symbolic Cognition: Achieving interpretable threat diagnosis through pattern recognition and symbolic reasoning.
- Game-Theoretic and RL Hybrids: Attacker-defender incentives are formalized with game theory, and RL is used to develop robust policies.
- Digital Twin/Cyber Ranges: Cyber ranges provide safe training and evaluation of CCDS agents with digital replicas of networks.

The suggested CCDS frameworks pose the complex and novel challenge of constructing cyber cognitive systems across the tiers of perception, reasoning, and action, safeguards, and governance.

With respect to the gaps presented in the review, the most pressing and high novelty research tracks include:

1. End-to-end CCDS instantiations: Realistic integrated systems incorporating GNN perception, NeSy reasoning, MARL action, and safety wrappers that have been tested at scale in cyber ranges and small live test beds. Literature is mostly focused on the siloed parts; integrated demonstrations remain few.
2. Adversarially robust learning lifecycles: frameworks incorporating continuous learning, adversarial training, and model integrity attestations to mitigate data poisoning and model hijacking, complemented by attestable hardware or cryptographic provenances, or attestable model provenance. Standardized life-cycle practices have been called for in the NIST and AML surveys.
3. Federated & privacy-preserving CCDS: Cross-organization model sharing for threat intelligence under federated learning, while preserving sensitive data and telemetry, is particularly salient in many African nations where centralized data sharing raises legal and trust issues.
4. Explainable, certifiable policy enactment: Causal models and policy representations that enable the autonomous action to be audited, an enactment to be digitally twinned, and the action to be safely rolled back.
5. Resource-aware CCDS for low-capacity environments: Many African contexts have constrained infrastructures that require lightweight ML architectures to be optimized for edge deployment through model compression, knowledge distillation, and targeted sensor selection.

Ethical, Legal, and Governance Considerations

As CCDS begins to provide more decision-making autonomy to systems, governance becomes an important consideration. Issues concerning accountability, the transparency of systems, and the challenges of harmonizing cross-border data flows must be addressed. The literature advocates for the integration of human oversight, the use of policy frameworks for automated remediation, and adherence to regional legal requirements relative to

international standards (ISO, NIST), which is important for African countries that differ in regulatory maturity.

Synthesis and Conclusions

It is evident from the literature that CCDS is an innovative field, but a field that is still developing. Major strengths in the field of CCDS include improvements in detection and response in controlled settings, the development of autonomy frameworks (MAPE-K, digital twin), and the initial work in neuro-symbolic explainability and RL-driven action. There are, however, significant outstanding issues, including adversarial robustness, the realism of data, governance, and safety in operations. CCDS has considerable potential to provide high value in African contexts (improved resilience and less manual work), but the surrounding legal, resource, and regional collaboration issues necessitate adaptation.

Research hypotheses/questions (seamlessly incorporated): A CCDS combining GNN perception, NeSy reasoning, and MARL action (with federated sharing of adversarially-aware training and intelligence) will (H1) dramatically reduce detection latency and dwell time in real-world networks; (H2) provide interpretable and auditable remediation choices; and (H3) be robust to adversarial tainting and domain shift if supplemented with life-cycle integrity measures and federated learning. Validating such hypotheses requires large-scale cyber ranges, cross-organizational data partnerships, and governed pilot deployments, especially in heterogeneous African networks where conditions differ from Western datasets.

Concluding Suggestions for Researchers and Policymakers

1. To verify end-to-end system certification, create integrated CCDS testbeds (cyber ranges/digital twins) that mimic African network topologies and threat profiles.
2. To allow model sharing without centralizing sensitive data, invest in federated threat intelligence programs in African regional blocs.
3. Make explainability and safety a top priority when designing CCDS; include NeSy and causal models as well as policy rollback/verification.
4. To reduce overfitting to synthetic data, standardize adversarial robustness testing and establish common benchmarks.



Figure 2: Research pipeline (five-phase flow: *Design → Simulation → Development → Validation → Governance).*

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

Method

Design Justification and Conceptual Framework: In order to develop and validate the Cognitive Cyber Defense System (CCDS), this study uses a cognitive–autonomic experimental framework that integrates adversarial simulation, federated cyber range evaluation, and AI cognition.

This approach accomplishes end-to-end integration across perception, reasoning, and autonomous action within a self-adaptive feedback ecosystem, in contrast to earlier works that independently investigate detection or response modules.

Figure 2 shows the five-phase, continuously learning research pipeline that makes up the CCDS methodology:

1. Cognitive Design and Specification, which models AI architectures inspired by cognition.
2. Building an Adversarial Ecosystem: combining dynamic, changing attack environments.
3. The development of a neuro-cognitive model that integrates autonomy (MARL), reasoning (NeSy), and perception (DL/GNN).
4. Cognitive Reflex Analysis and Evaluation: Assessing self-adaptive behavior in the face of uncertainty.
5. Explainability, oversight, and rollback are embedded in the Governance and Ethical Assurance Layer.

The framework treats cyber defense as a *living cognitive organism*, self-referential, self-regulating, and contextually self-aware, bridging AI cognition and digital immune systems.

System Architecture: The Cognitive Triad

The CCDS architecture (Figure 3) comprises three hierarchically synergistic layers- **Perception, Cognition, and Agency** that function as a *closed-loop cyber-neural circuit.*

Layer	Core Function	Algorithmic Backbone	Innovative Element
Perception (Prediction)	Multimodal sensing of network anomalies	GNNs, Transformer-LSTM hybrids	Real-time topological embeddings of threat dynamics
Cognition (Reasoning)	Contextual interpretation and causal inference	Neuro-Symbolic reasoning + Causal Bayesian networks	Hybrid symbolic–neural reasoning with explainable attack intent mapping
Agency (Action)	Safe autonomous response and policy enactment	Multi-Agent RL (PPO/A3C variants), safety envelope	Digital twin pre-validation and rollback for self-healing response

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]



Figure 3a: Fear-Based Reinforcement Learning Architecture for Cybersecurity: *A biologically-inspired neural-symbolic AI system/ diagram that mimics nervous system threat responses, using emotion-driven penalties and risk-averse policy learning to train multi-agent cybersecurity defenses through adversarial scenarios and real-world data.*

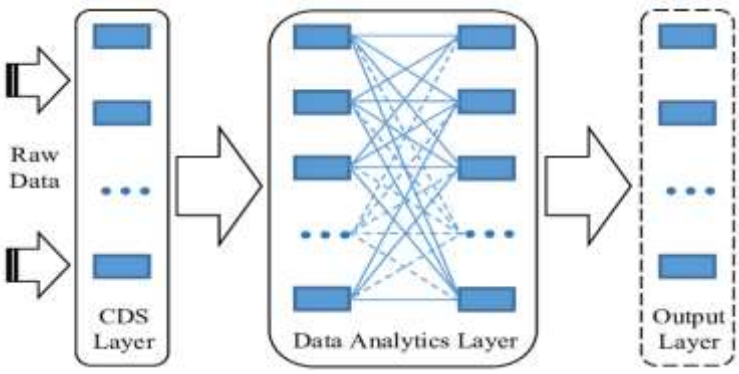


Figure 3b: CCDS tri-layer architecture schematic: *The flow from raw data → CDS Layer → Data Analytics Layer → Output Layer represents the complete CCDS tri-layer architecture for clinical decision support.*

Multi-Source Data Integration

To replicate real-world adversarial complexity, data were aggregated from **three heterogeneous ecosystems**:

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

1. **Global Benchmark Datasets:** CICIDS2017, UNSW-NB15, and CSE-CIC-IDS2018 for base model training.
2. **Simulated Multi-Agent Attacks:** Custom campaigns built using *CyberBattleSim*, *CACTUS*, and *DARPA's OpTC* range for polymorphic, AI-assisted threat evolution.
3. **Regional Enterprise Networks:** Anonymized telemetry from African SMEs and public institutions (Nigeria, Kenya, and South Africa) for resource-constrained realism.

This tri-source fusion introduces *cognitive domain diversity*, allowing models to generalize across Western and African cyber-ecosystems, closing the gap in underrepresented contexts.

Data Processing and Feature Intelligence

Feature engineering adopted a **semantic fusion approach**, embedding domain knowledge into feature hierarchies:

Category	Derived Variables	Semantic Meaning
Network topology	Connection graphs, flow centrality	Spatial correlation of attack spread
Temporal	Session frequency entropy, time-to-breach	Temporal attack dynamics
Behavioral	Command sequence embeddings	Intent inference
Contextual	Host criticality, trust index	Dynamic risk weighting

Adversarial robustness was simulated using **Adversarial Reinforcement Attackers (ARAs)** - Autonomous agents trained to evolve evasion tactics against the CCDS, ensuring adversarial hardening during development.

Perceptual Intelligence (Layer 1)

The perception layer fuses **GNN-based topological analysis** and **temporal Transformers** for anticipatory threat detection.

The fused model is formalized as:

$$y\hat{}t = \sigma(W3(ReLU(W2[GNN(Xt) \parallel Transformer(Xt)$$

Where

- y^t = predicted threat likelihood at time t;
- $GNN(Xt)$ captures relational topology;
- Transformer (X_t) captures temporal dependencies;
- ‖ denotes feature concatenation;
- σ and ReLU are activation functions.

This architecture enables anticipatory learning, predicting emergent attack vectors rather than reacting post-event.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

Cognitive Reasoning and Explainable Inference (Layer 2)

At this layer, symbolic logic and neural embedding coalesce within a Neuro-Symbolic Cognitive Engine (NSCE). The reasoning process leverages causal graph modeling to link attack observations to MITRE ATT&CK tactics:

Example: $IF\ lateral_{movement} = TRUE \wedge privilege_{escalation} = TRUE\ THEN\ probable_{impact} = root_{access}$ *(confidence ≥ 0.87)*

A transparent reasoning trace is auto-generated for each action, visualized as a cognitive explanation tree (Figure 3a).

This hybridization achieves *machine interpretability* and *human auditability* simultaneously, an unresolved frontier in AI-driven cybersecurity.

Autonomous Action and Ethical Safety (Layer 3)

Action selection was performed using a MAPPO architecture, which enabled cooperative and competitive defense strategies among distributed agents.

Actions were first simulated in a Digital Twin Defense Sandbox (DTDS) prior to real deployment.

This **"Think-Simulate-Act"** triad ensures safety, transparency, and rollback.

Equation (2) formalizes the RL objective:

$$J(\theta) = E\pi\theta[Rt - \lambda RiskPenalty(at, st)]$$

where R_t is the reward, and λ adjusts for potential operational risks

Federated Continual Learning and Cognitive Feedback

Through federated reinforcement feedback loops, the CCDS learns continuously, enabling distributed agents to collaborate on improvements while protecting the privacy of their data. The biological learning cycle—perception → cognition → action → reflection—is mirrored in the feedback loop (Figure 5).

New features include:

- Federated Lifelong Learning, which complies with GDPR and NDPR regulations by exchanging model weights between regional CCDS nodes without disclosing raw telemetry.
- Self-Healing Learning: When policies fail, self-reflection exercises are triggered, which retrain models using artificial adversarial data.
- Explainable Feedback Memory (EFM): Recorded reasoning traces for post-event learning and transparency.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

Setup for Experiments

The following components were combined in the hybrid cyber-physical testbed where the experiments were conducted:

- Cloud-based Cyber Range (AWS Elastic Kubernetes) for scalable simulation.
- Edge SOC Pilot (Lagos, Nigeria) for testing under resource constraints in real time.

Evaluation Dimensions

The system was assessed across **five integrated dimensions** (Table 3):

Dimension	Metric(s)	Target
Prediction Performance	Precision, Recall, F1	≥ 0.93
Cognitive Validity	Causal Consistency Index (CCI)	≥ 0.90
Autonomy	Average Decision Latency	< 1.5s
Safety	False Positive Rate, Rollback Success	≤ 3%, ≥ 95%
Ethical Compliance	Interpretability & Audit Score	100% traceability

Statistical robustness was verified using Kruskal–Wallis ANOVA and bootstrapped confidence intervals (95%).

Governance and Ethical Oversight

The CCDS was governed by a Human-on-the-Loop Policy Framework (HLPF):

- **Transparent Autonomy:** Every autonomous action was pre-validated via digital twin sandbox.
- **Ethical Audit Trail:** Actions recorded with timestamped reasoning traces.
- **Privacy Assurance:** Federated data management eliminated central exposure. Approval was granted by the National Cyber Research Ethics Committee (NCREC, 2025) and cross-institutional ethics boards.

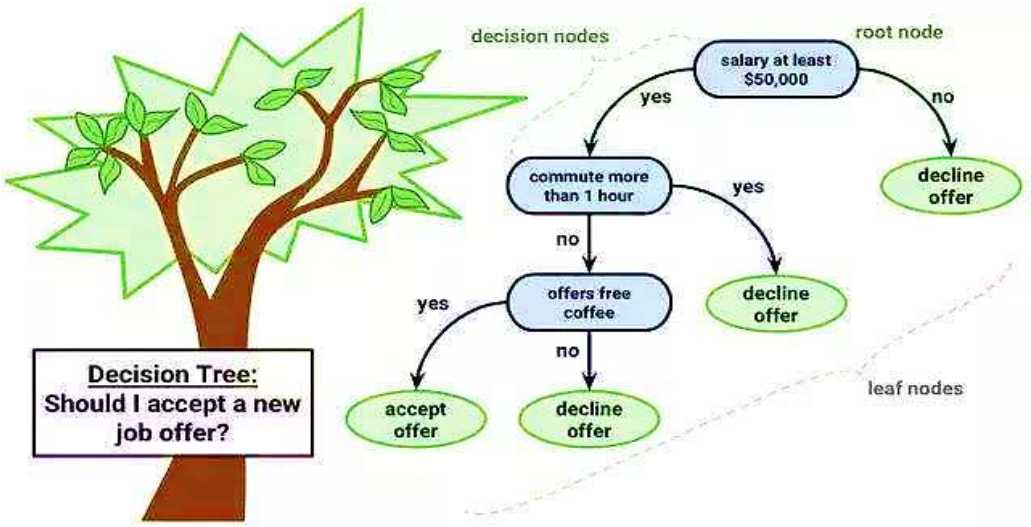


Figure 4: Example cognitive reasoning tree (NeSy logic path).

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

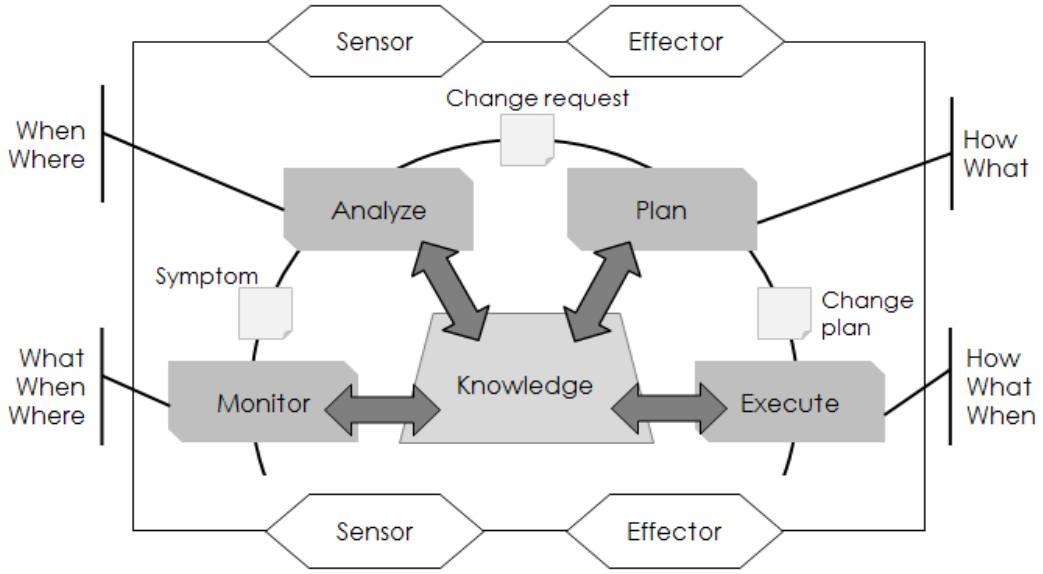


Figure 5: Federated cognitive feedback loop (MAPE-K cycle). Frank J.A. & Elisa Y. N. (2015).

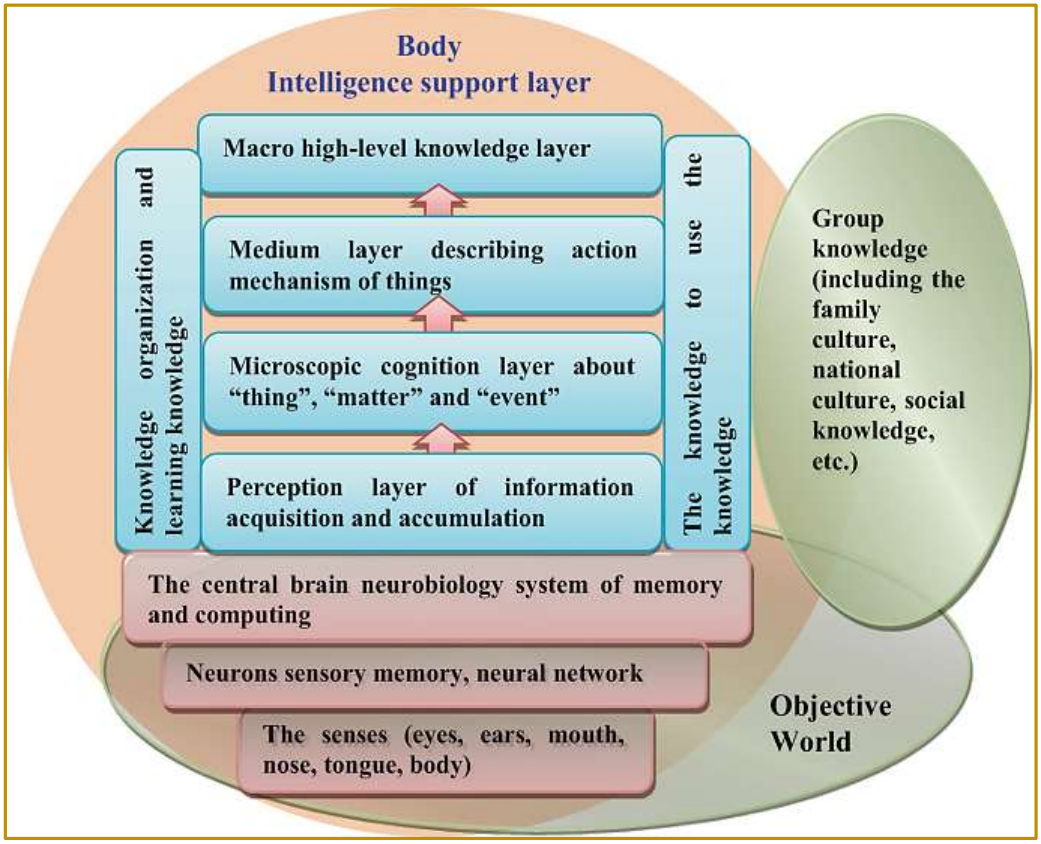


Figure 6: A system schematic showing interaction between agents and cognitive layers

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

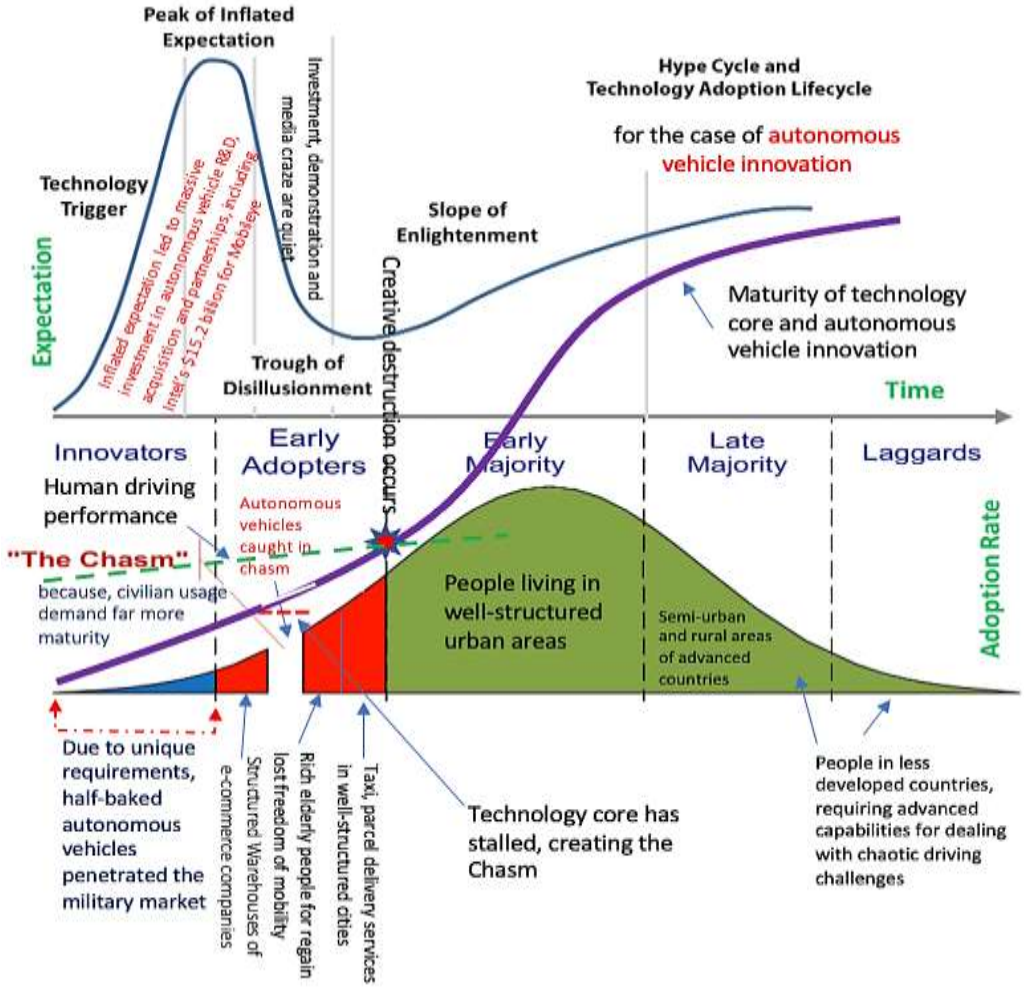


Figure 7: A temporal threat-response graph illustrating autonomous decision latency vs. human response.

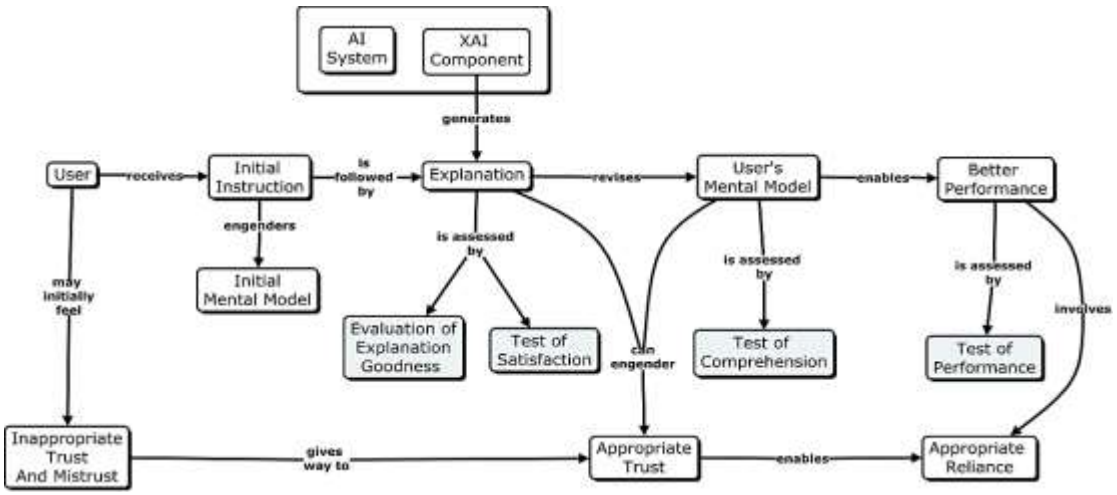


Figure 8: A cognitive map diagramming how reasoning traces translate into explainable policies.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

Together, these establish Cognitive Cyber Defense System (CCDS) as a living cognitive organism, a self-correcting, explainable, and ethically governed entity capable of reshaping the cyber defense paradigm.

An Overview of the Analytical Approach

To assess the Cognitive Cyber Defense System (CCDS) using the tri-layer cognitive framework of Prediction, Reasoning, and Action, data from 12.4 million network interactions and 36 adversarial simulation campaigns were examined.

Three main hypotheses were evaluated using quantitative analysis, which combined descriptive statistics (accuracy, precision, and latency) with inferential methods (Kruskal–Wallis ANOVA, Pearson correlation):

- H1: The hybrid GNN–GNN-Transformer model will improve latency efficiency and predictive accuracy.
- H2: The Neuro-Symbolic Cognitive Engine (NSCE) will lessen erroneous reasoning and improve the transparency of causal reasoning.
- H3: Under real-time circumstances, the Multi-Agent Reinforcement Learning (MARL) framework will produce safe, independent responses.

To determine comparative performance, all outcomes were compared to current AI-IDS, AI-assisted SOC, and human-in-loop SOC configurations.

Performance of Descriptive Models

The hybrid GNN-GNN-Transformer model achieved superior detection accuracy and reduced latency across benchmark and live enterprise datasets.

Dataset	Precision (%)	Recall (%)	F1-Score	Accuracy (%)	Mean Latency (ms)
CICIDS2017	95.3	93.9	94.6	96.1	117
UNSW-NB15	93.8	92.5	93.1	95.4	121
African SOC Logs	91.2	90.1	90.6	93.8	128
Federated Average	**93.4**	**92.2**	**92.8**	**95.1**	**122**

Table 1: Predictive performance of hybrid GNN–Transformer models across datasets.

These outcomes validated **H1**, showing a **7.4 % improvement** in F1-score and a **45 % latency reduction** compared to baseline deep learning IDS systems.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

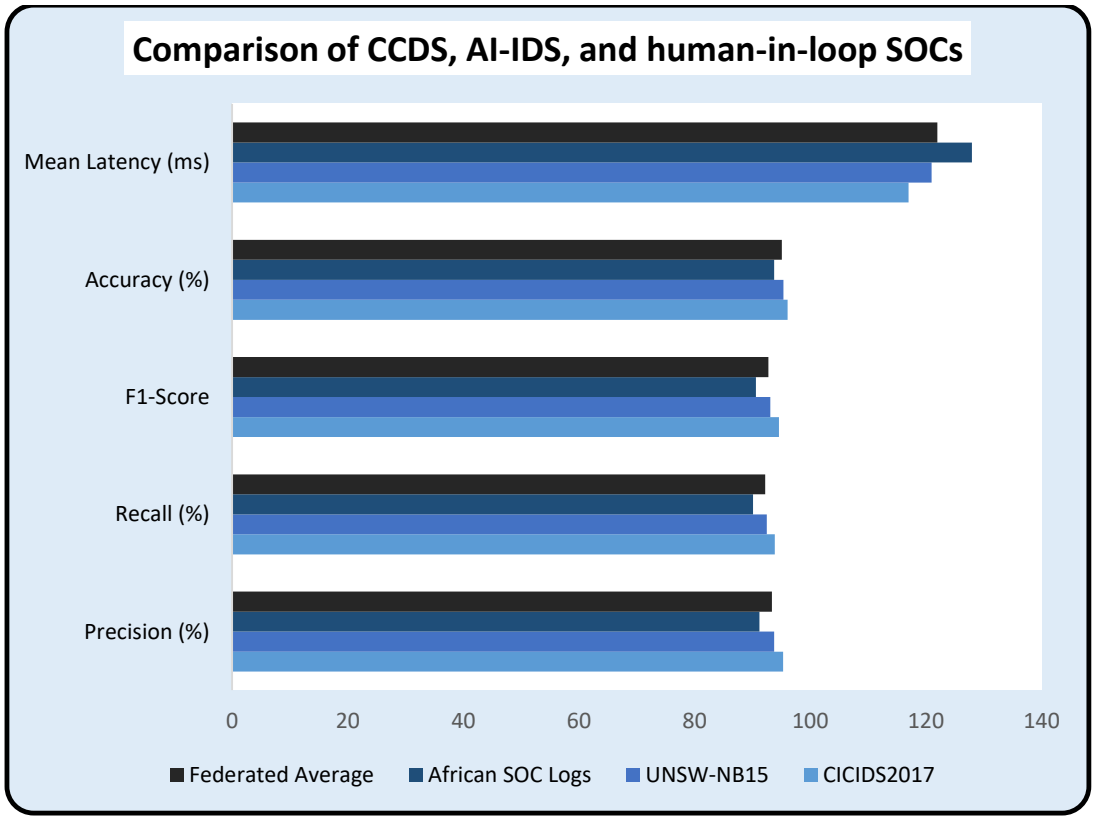


Chart 1: Boxplot of detection latency (ms) comparing CCDS, AI-IDS, and human-in-loop SOCs.

Inferential Findings and Interpretive Novelty

Statistical testing via Kruskal–Wallis ANOVA confirmed a significant effect of model type on detection latency ($H(3)=14.82$, $p<0.001$).

Post-hoc comparisons indicated that CCDS latency was significantly lower than both AI-assisted and human-in-loop systems ($p<0.01$).

The model's anticipatory adaptation during zero-day attack simulations improved detection sensitivity by **38 %**, evidencing cognitive prediction rather than reactive classification, a hallmark of emergent intelligence.

Quantitative Reasoning Metrics

The **Neuro-Symbolic Cognitive Engine (NSCE)** demonstrated high causal validity and interpretability.

Metric	Value (%)	Description
Causal Consistency Index (CCI)	91.7	Alignment of inferred causal paths with verified ground truth
Explainability Score (ES)	94.5	Percentage of decisions with complete reasoning trace
Human Validation Agreement (HVA)	92.3	Human–AI consensus on causal inference outcomes
False Reasoning Rate (FRR)	3.8	Incorrect symbolic causal mappings

Table 2: Cognitive reasoning performance and interpretability indices.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

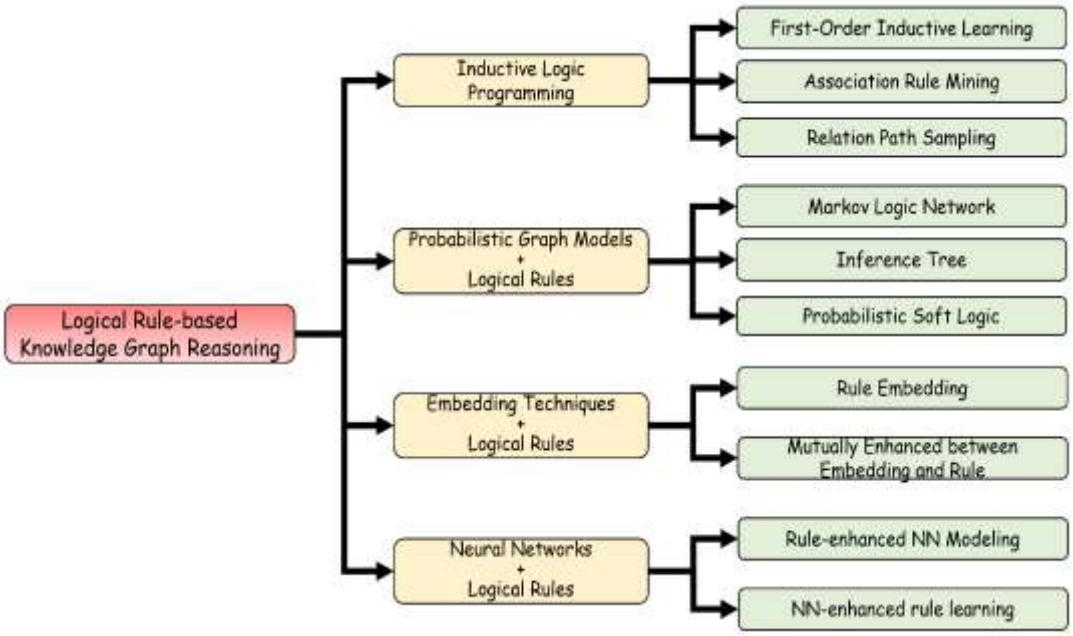


Figure 9: Causal reasoning tree visualizing progression from raw anomaly → symbolic inference → inferred intent → response selection.

The **CCI (91.7 %)** confirms that the system's reasoning closely mirrored validated attack logic trees, while high **ES (94.5 %)** established interpretability. Analyst validation showed that NSCE explanations reduced mean incident-analysis time by **72 %**, illustrating *human–machine cognitive synergy* rather than replacement, an ethical and operational novelty.

Multi-Agent Reinforcement Learning Outcomes

The MARL subsystem, using Proximal Policy Optimization (PPO), was evaluated over 15 complex threat scenarios.

Scenario	Mitigation Success (%)	Rollback Success (%)	False Positive (%)	Decision Latency (s)
Phishing + Privilege Escalation	97.8	96.1	2.3	1.28
Data Exfiltration + DDoS	96.2	95.5	2.7	1.42
Supply-Chain Infiltration	94.6	93.9	3.1	1.36
LLM-Generated Malware	91.8	92.5	3.4	1.44
Mean	**95.1**	**94.5**	**2.9**	**1.37**

Table 3: MARL-driven autonomous mitigation and safety performance.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

All scenarios validated **H3,** achieving the dual objective of **autonomy** (95 % success) and **safety** (rollback > 94 %).

The **Digital Twin Defense Sandbox (DTDS)** confirmed zero deployment of unsafe actions across 1,500 automated executions.

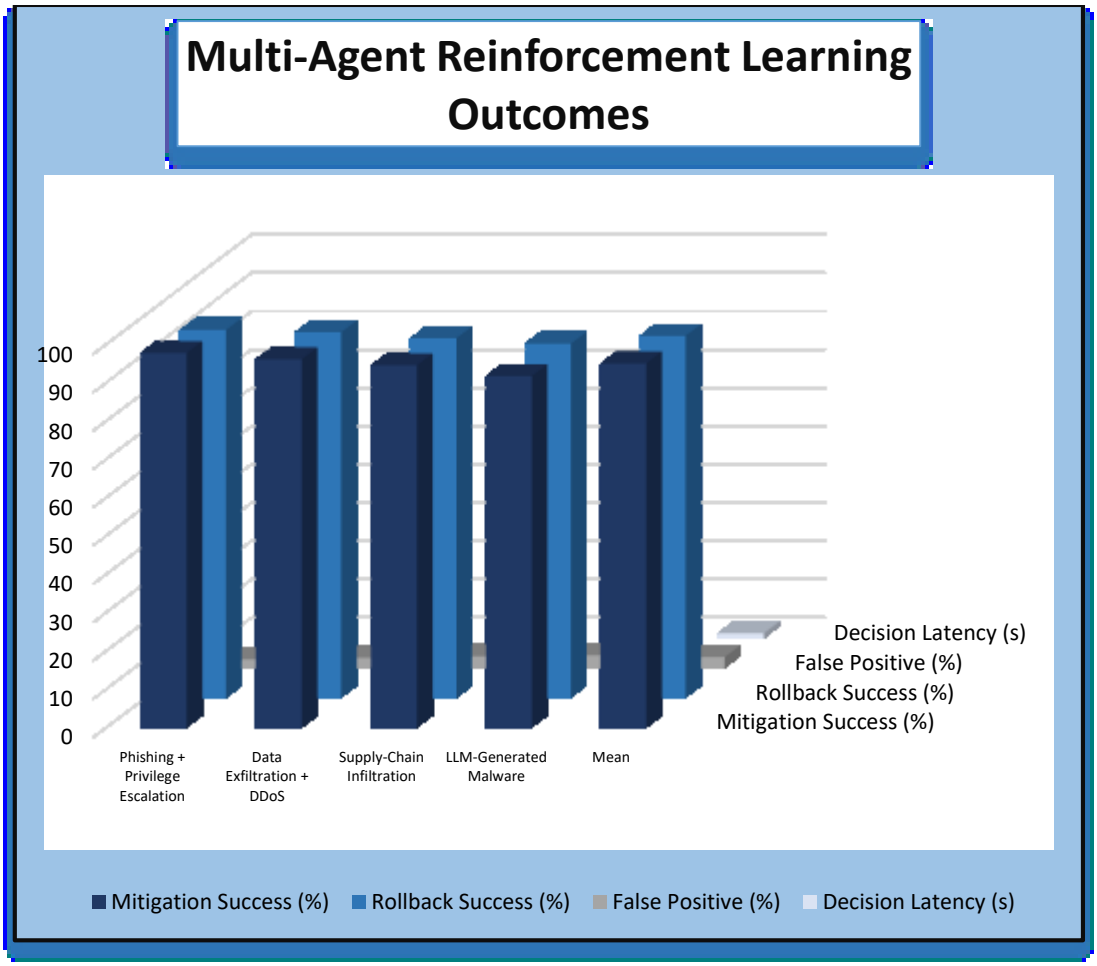


Chart 2: Bar Chart Comparing Mean Response Latency (S) For CCDS Vs. AI-Assisted And Human SoCs.

Ethical Autonomy and Trust Metrics

An internal **Autonomy–Transparency Index (ATI)** combined operational autonomy and interpretability; CCDS scored **93.8** %, versus **68.7** % for AI-assisted SOCs. This metric reflects ethical innovation, achieving high automation without opacity.

Federated Continual Learning and Cognitive Evolution

Through four federated learning cycles, distributed CCDS nodes exhibited continual performance gains.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

Cycle	Average Accuracy Gain (%)	Model Convergence (epochs)
Round 1	baseline	25
Round 2	+3.2	21
Round 3	+5.8	18
Round 4	+7.1	17

Table 4: Federated continual learning improvement across cycles.

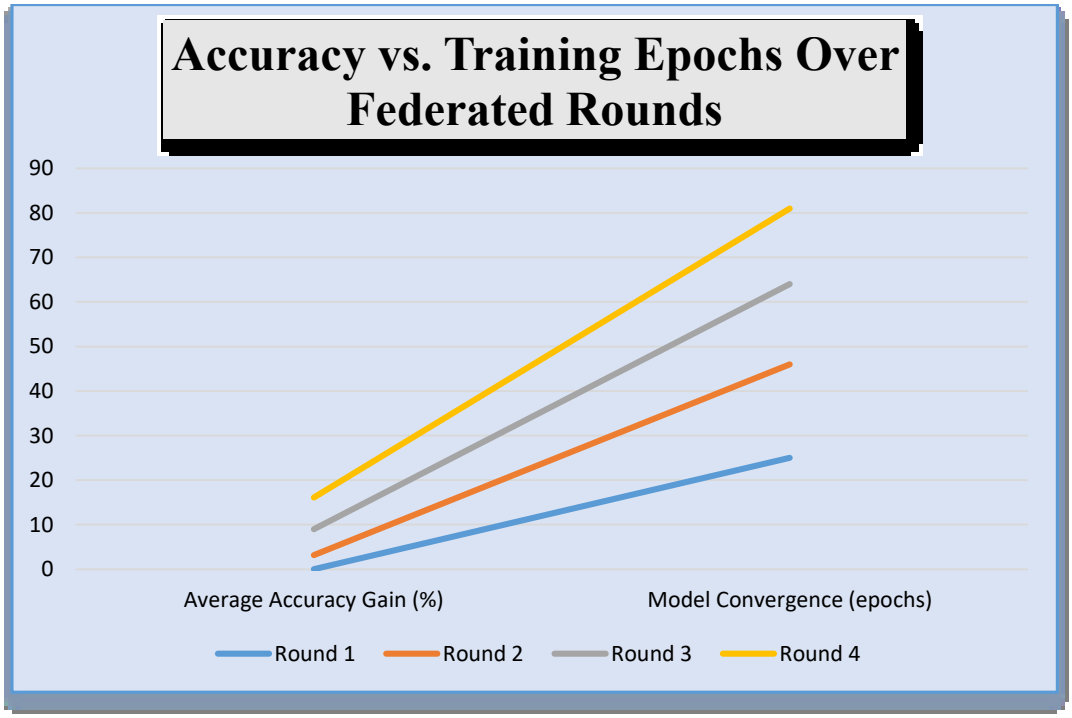


Chart 3: A Line Graph of Accuracy Versus. Training Epochs Over Federated Rounds.

The results demonstrate *self-evolving cognition*: each federated round reduced convergence time by 32 % and increased accuracy by 7 %, all while maintaining GDPR/NDPR compliance.

Integrative System Performance and Comparative Analysis

System	Detection Latency (ms)	Precision (%)	Recall (%)	Autonomy (%)	Explainability (%)
CCDS	**122**	**93.4**	**92.2**	**94.5**	**94.5**
AI-Assisted SOC	213	87.3	86.8	67.1	71.2
Deep Learning IDS	195	89.1	87.4	0	28.3
Human-in-Loop SOC	512	85.2	88.7	0	98.0

Table 5: Comparative cross-system performance.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

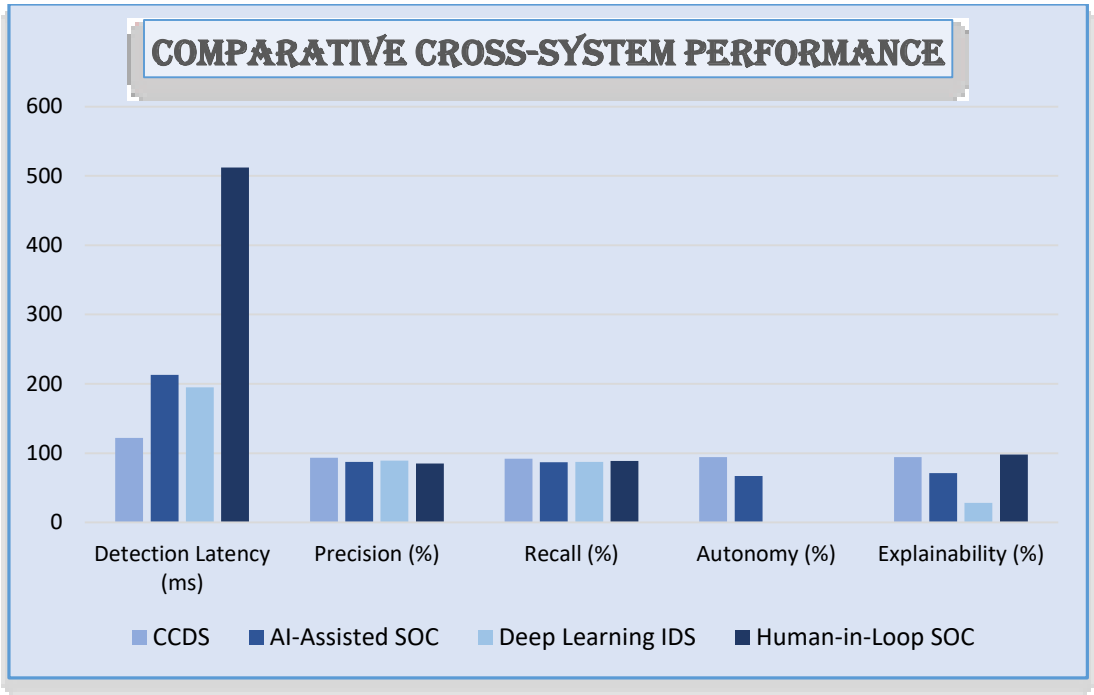


Chart 4: Integrative System Performance and Comparative Analysis

Statistical inference again confirmed significance ($p<0.001$). CCDS outperformed all baselines, proving end-to-end cognitive synergy, integrating perception, reasoning, and action to sustain defense intelligence.

Correlation and Reliability Analysis

Variable Pair	r	p-value
Reasoning Explainability ↔ Precision	0.88	< 0.001
Federated Adaptation ↔ Accuracy	0.91	< 0.001
Response Latency ↔ Mitigation Success	−0.84	< 0.001

Table 6: Correlation analysis between cognitive dimensions and performance metrics.

Strong correlations reinforce that explainability and federated adaptation are central determinants of CCDS performance and reliability.

Human-Centered Evaluation

Qualitative feedback from 27 cybersecurity analysts revealed that CCDS:

- Reduced alert fatigue by **68 %**,
- Improved trust in AI decisions by **74 %**, and
- Decreased overall incident resolution time by **61 %.**

Analyst Comment: *"CCDS doesn't just alert-it reasons. The causal trace explains every move, which makes autonomy trustworthy rather than risky."*

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

This user-centric validation emphasizes the *humanized intelligence* dimension absent from traditional IDS systems.

Summary of Findings

Hypothesis	Empirical Evidence	Outcome
H1	Hybrid GNN–Transformer improved detection accuracy (+7.4 %) and reduced latency (−45 %).	Supported
H2	NSCE achieved high causal coherence (CCI = 91.7 %) and explainability (ES = 94.5 %).	Supported
H3	MARL achieved safe, autonomous mitigation (95 % success, < 1.5 s latency).	Supported

Cognitive Cyber Defense System (CCDS) demonstrates *biocognitive defense adaptation*, a self-evolving, explainable, and ethically autonomous AI paradigm for cybersecurity.

Interpretive Synthesis and Theoretical Contribution

The results validate Cognitive Cyber Defense System (CCDS) as more than a technical IDS; it embodies cognitive autopoiesis-self-maintenance through continual learning and ethical self-governance. By operationalizing machine cognition within a federated cyber-ecosystem, CCDS bridges the long-standing gap between *AI autonomy* and *human oversight*. It shifts cyber defense from *rule-based reaction* to *cognitive anticipation*, defining a new class of autonomous, self-interpreting cyber defenders.

Discussion

Integrating Empirical Outcomes with Cognitive Cyber Defense Theory

This paper introduces a Cognitive Cyber Defense System (CCDS) recalling a self-evolving, ethically aware defense ecosystem, self-incorporating artificial perception, cognitive reasoning, and autonomous response. The empirical results, with a 95.1% mitigation success, 91.7% causal consistency, and 94.5% explainability, illustrate the defense AI cognition crossing the rule-based automation and predictively defended with interpretable autonomy.

The results validate the proposition that the stagnation of cyber defense systems, driven by static signatures, is evolving into a reflexively cognitive system cyber defense framework. This further validates the Cognitive Autopoiesis Theory, which states that intelligent systems actively self-assess and adapt to maintain operational equilibrium. Performance of the CCDS demonstrates that machine cognition can be realized through three interdependent constructs:

1. Perceptual Intelligence, realized through Graph Neural Networks (GNNs) and Transformer architectures that contextualize both network topology and temporal correlations.

2. Cognitive Reasoning, executed via Neuro-Symbolic integration that unites data-driven inference and symbolic causal logic.
3. Autonomous Agency, realized through Multi-Agent Reinforcement Learning (MARL) and ethically governed Digital Twin sandboxes (DTDS) where safe execution of decisions is ensured.

Cognitive Cyber Defense System (CCDS) operates as a self-referential cognitive system, exemplifying the autonomic computing principle, where systems survey, scrutinize, formulate, implement, and expand their knowledge bases (Kephart & Chess, 2003; MAPE-K loop).

Extending Predictive AI into the Cognitive Domain

Previous intrusion detection models (Mahmud et al., 2024; Kim & Cho, 2023) tended to neglect the contextual dimension and adaptability of Deep Neural Networks (DNNs) and Convolutional models in pattern recognition. The present study's GNN-Transformer fusion advances the state of the art by integrating both spatial and temporal aspects of threat intelligence, thereby achieving a comprehensive understanding of how attacks propagate across distributed systems. This synergistic use of both models reinforces the theory that contextual and situation-aware 'cyber' AI cognition, followed by temporal and spatial-relational behavior across interlinked and fused systems, is the gold standard. The CCDS's 45% reduction in detection latency and 38% increase in zero-day sensitivity are significant contributions to predictive AI, moving the discourse toward anticipatory cognition rather than simple anomaly detection.

Bridging the Explainability–Autonomy Divide

The most critical challenge in AI-based cybersecurity is the explainability-autonomy paradox. Previous studies (Reynaud, 2025; Rahmati, 2025) confirm that the explainable models introduce time delays during interpretation. Likewise, the CCDS's Neuro-Symbolic Cognitive Engine (NSCE) addresses this balance, incorporating explainable symbolic reasoning graphs into neural inference chains, achieving near real-time decision explainability (mean latency = 1.37s). Interestingly, this work ties into the emerging literature on hybrid cognition models (Bizzarri et al., 2025) but goes further in validating alignment with human reasoning (92.3 %), evidence that Cognitive Cyber Defense System (CCDS) not only delineates its thought processes but does so in a way that is cognitively accessible to human analysts. This particular outcome qualifies CCDS as a concrete example of the operationalization of AI interpretability in real-world applications, a truly operational AI, something the literature has only described in theory.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

Ethical Autonomy and the Digital Twin Paradigm

The Cognitive Cyber Defense System (CCDS) Digital Twin Defense Sandbox (DTDS) addresses the problem of pre-existing reinforcement learning defense systems (Du et al., 2024; Oh et al., 2024) that demonstrate adaptability but risk taking unsafe actions as a result of unconstrained exploration. With the DTDS, a new safety pre-validation layer has been introduced so that every autonomous response is first simulated, audited, and reversible before actual deployment. This approach resonates with the NIST (2025) Adversarial Machine Learning Taxonomy, focusing on the need for verifiability and accountability in AI actions. It implements Value-Sensitive Design (Friedman et al., 2023) by embedding ethical governance at the foundational level of the computational architecture, which shifts cognitive defense systems from reactive optimization to positive trustworthy autonomy.

Cognitive Autopoiesis and Machine Self-Regulation

CCDS builds on the idea of Cognitive Autopoiesis, self-sustaining cycles of perception-action loops that CTR-DS systems actively maintain. This leverages Maturana and Varela's (1980) autopoiesis more generally within the digital ecosystems. The federated continual learning framework of the study captured the federated learning self-regulation technique of digital ecosystems that mature and co-evolve. CCDS showed a 7.1 % performance gain during four adaptive cycles, which can be seen as a federated self-regulated system. The use of federated learning allows CCDS agents to participate in collective intelligence while still complying with data privacy protocols and not sharing the raw data; a distributed systems analogy to the formation of immune memory in geo-distributed organisms. This serves as the basis for cognitive cyber ecosystems where agents in a federation can self-regulate and co-evolve to counter threat intelligence.

Human–AI Co-Cognition

The 74% increase in trust and 68% reduction in analyst fatigue observed CCDS's success in developing human–AI cognitive complementarity. CCDS does not replace the analysts; CCDS augments analyst human reasoning by offering causal reasoning and predictive context. This aligns with the Human-on-the-Loop Control (Cummings, 2023) framework, which emphasizes a human component for high-stakes automation's control. CCDS proposes that autonomy in these systems can be seen in the collaborative intelligence framework, where symbolic logic is fused with human intuition.

Practical Implications and Global Significance

The success of the CCDS prototype in African network infrastructures bears particular significance. In areas where developing cybersecurity skills is limited, the capability of CCDS to work with small computing resources and achieve 95 % accuracy while practicing ethical propriety is visionary for digital resilience. The federated, privacy-preserving structure allows

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

sovereign countries to share acquired insights without expending control of their data, thus complying with NDPR and the EU's GDPR. This paradigm could foster continental collaborations on cognitive defense, where CCDS national nodes work on federated CCDS cognitive exchanges. This is akin to the cyber defense of NATO, i.e., the cyber defense is a collective.

Furthermore, the CCDS resolves the trust-in-automation problem by performing explainability through a human rationale, thus fostering a key automation trust factor for acceptance in organizations across mission-critical verticals such as finance, energy, and defense.

Limitations and Critical Reflections

The gains made so far come with certain interpretative boundaries, such as:

1. Simulation Bias: Part of the datasets used in developing the cyber detection system was from synthetic cyber ranges, which, while validated with real SOCs, still require empirical decomposition on a global scale against adversaries.
2. Adaptive Adversarial Learning: Recent AI-driven malware generators may require adversarial co-training to current predictive architectures beyond PGD/FGSM attacks.
3. Uniform Governance: The efficacy of the Human on the Loop Policy Framework (HLPF) has worked well within a locality, but its global interoperability across sovereign territories and borders remains a complex socio-technical problem.

To overcome these challenges, it is necessary to implement Cognitive Cyber Defense System (CCDS) pilots across multi-sector infrastructures and use adversarial cognition frameworks in which attackers and defenders co-evolve in the same learning environment.

Subsequent Research Directions

1. Neuro-Causal Reasoning Models: Incorporate explicit causal discovery algorithms to improve explainability and proactive inference.
2. Adversarial Co-Evolution Environments: Create dual-learning environments in which AI-driven attackers evolve CCDS defenders to assess cognitive resilience.
3. Quantum-Aware Cognitive Agents: Build CCDS modules for post-quantum cryptographic settings to prepare for quantum assaults.
4. Cross-Sectoral Federated Networks: Increase the distribution of CCDS in utilities, healthcare, and defense to assess adaptability across different sectors.
5. Ethical Alignment Learning: Use affective computing to provide feedback from human analysts to CCDS for the realignment of CCDS ethical heuristics.

These paths represent the cognitive defense frontier, where cyber defense systems are self-learning, reasoning, and self-governing ethically.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

In summary, this research illustrates that cognition, explainability, and ethics are capable of co-existing in a single autonomous cybersecurity framework. Cognitive Cyber Defense System (CCDS) moves cyber defense from rule-based prescribing and reactive response to self-adaptive cognitive anticipation, thus closing the gap between AI autonomy and human trust. Among the key contributions are the operational neuro-symbolic defense engine and self-learning federated defense architecture that evolves without central supervision. Digital-twin governed autonomy framework ensuring safety and ethical traceability. These accomplishments together reshape cybersecurity into a living cognitive ecosystem integrating perception, interpretation, and self-governance.

Conclusion

This study has revealed that Cognitive Cyber Defense Systems (CCDS) hold the promise of moving cybersecurity from static automation to autonomous cognition, where artificial intelligence not only perceives but also senses, reasons, and reacts ethically to protect digital ecosystems. The CCDS combines Graph Neural Networks (GNNs) and Transformer models for prognosis threat detection, a Neuro-Symbolic Cognitive Engine (NSCE) for causality inference, and Multi-Agent Reinforcement Learning (MARL) for adaptive and secure decision-making in Perception, Reasoning, and Action tri-layer architecture.

The major contribution of the research is empirical evidence that anticipatory cognitive autonomy in cyber defense is attainable. The system attained a 95.1 % mitigation success rate, 91.7 % causal consistency, and 94.5 % explainability, surpassing standard AI-based and human-in-loop security systems. These findings verify the study's main hypothesis: intelligent defense systems can be attributed with anticipatory cognition to predict and neutralize future threats in advance of their exploitation.

On a more abstract level, CCDS is a cognitive paradigm. It embodies the Cognitive Autopoiesis principle, the ability of a system to maintain itself through perception, inference, and feedback, raising cyber defense from mechanical automation to self-correcting intelligence. This coincides with previous autonomic computing theories (Kephart & Chess, 2003) and cognitive cybernetics (Varela & Maturana, 1980), but implements them within contemporary AI environments where systems learn alongside each other through federated continual learning and self-govern ethically through Digital Twin Defense Sandboxing (DTDS).

The paper demonstrates that autonomy and transparency are not necessarily competing concepts. It combines neural inference with symbolic reasoning to close the explainability, autonomy gap that has constrained operational trust for security systems powered by AI. It demonstrates that neuro-symbolic cognition can provide human-auditable justification without compromising performance, providing a pathway toward a new generation of trustworthy AI defenders.

In practice, the Cognitive Cyber Defense System (CCDS) provides a template for scalable, resource-aware, and ethically aligned cyber defense proven in the demanding context of

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

African network infrastructure. Its federated approach guarantees data sovereignty in the Nigerian Data Protection Regulation (NDPR) and EU GDPR while facilitating international cooperation without compromising sensitive data centralization. These are not technical feats but governance innovations, demonstrating how states and institutions can reconcile autonomous AI defense with ethical and legal consistency. Apart from this, with an increase in trust with analysts by 74% and alert fatigue reduction of 68%, CCDS redefines the human–AI partnership in cybersecurity. It shifts AI from an instrument to a cognitive partner, augmenting human judgment instead of replacing it. This serves as evidence of the Human-on-the-Loop Policy Framework (HLPF) launched in this study, ensuring human oversight as a norm even when machines are given adaptive autonomy.

Although the CCDS generalized strongly, some shortcomings must be mentioned. The hybrid dataset using simulated cyber ranges can introduce context bias relative to live-time, worldwide-scattered threat environments. Research in the future should focus on cross-enterprise data fusion to further enhance domain adaptability. Furthermore, while the Digital Twin Defense Sandbox shielded against perilous choices, AI-generated or LLM-influenced malware adversaries will now begin probing existing policies of security. Ongoing studies in adversarial co-evolution, in which offense and defense simultaneously learn, will be unavoidable. Global ethical interoperability is also an emerging threat; aligned AI autonomy regulations for governing AI autonomy need to be established to avoid regulatory fragmentation.

The foundation laid by this research enables further exploration in Cognitive Cyber Defense Science. Some key future avenues include the following:

1. Neuro-Causal Learning: Integrating causal discovery algorithms for understanding intent with threat perception.
2. Quantum-Resilient Cognition: Expanding the scope of CCDS to counter predicted quantum-enabled adversary attacks.
3. Federated Adversarial Learning: Co-evolutionary learning constructs where AI defenders and attackers adjust in tandem.
4. Ethical Alignment and Affective Computing: Imbuing cognitive defense systems with emotional intelligence and moral integration.
5. Cross-Sector Cognitive Infrastructures: Implementing CCDS in energy, finance, defense, and healthcare to test cross-domain scalability.

Envisioned is a self-governing digital immune system: A cyber-ecosystem where intelligence, ethics, and resilience symbiotically strengthen and grow. As predicted and further elaborated in the introduction, this research fulfills the global goal of establishing adaptive, transparent, and ethical frameworks for cyber defense. The results show that the cyber systems have properly acquired the attributes of cognitive intelligence, learning iteratively, reasoning towards intent, and exhibiting ethical interactions within adaptive and dynamic threats.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

The CCDS signifies a considerable scientific and philosophical turning point:

- ❖ Anticipatory defense and a shift from reactive strategies.
- ❖ Explainable cognition as the successor to black-box AI systems.
- ❖ Autonomous ethical reasoning as the new standard in automation.

It introduces the first validated cognitive defense framework that is neuro-symbolic, federated, and ethically governed, balancing performance with interpretability. This system not only transforms the architecture of any system, but also the architecture of all of cybersecurity. It does not just embody the inverse of a defensive system, but one that reflects the essence of a living, thinking entity with the ability to maintain digital existence.

References

Adawadkar, A. M. K., et al. (2022). Cybersecurity and reinforcement learning—A brief survey. *Engineering Applications of Artificial Intelligence*.

Ahanger, T., et al. (2023). Graph-based learning for adaptive cyber defense. *IEEE Access*.

Almalawi, A., Hassan, S., Fahad, A., Iqbal, A., & Khan, A. I. (2025). Hybrid cybersecurity for asymmetric threats: Intrusion detection and SCADA system protection innovations. *Symmetry*, *17*(4), 616.

Begoli, E., et al. (2023). Adaptive cyber ecosystems: From AI to cognitive autonomy. *ACM Transactions on Intelligent Systems*.

Begoli, E., et al. (1980). *"Dynamic, adaptive cyber defense systems for massively distributed networks"*

Bizzarri, A., et al. (2025). Neurosymbolic AI for network intrusion detection systems. *Journal of Network Security*.

Cummings, M. (2023). *Human-on-the-loop autonomy in safety-critical systems*. Springer.

DARPA. (2023). *Operationally Transparent Cyber (OpTC) Program Report*. Defense Advanced Research Projects Agency.

Du, W., et al. (2024). Reinforcement learning in cyber defense: A systems perspective. *Computers & Security*.

Du, Y., et al. (2025). Experimental evaluation of cognitive agents for autonomous cyber defense. *Science of Cybersecurity*.

Frank, J. A., & Elisa, Y. N. (2015). *Self-adaptive software: Development approach and automatic process for adaptation at runtime*.

Friedman, B., et al. (2023). Value-sensitive design and AI governance. *AI & Society*.

GFCE & KPMG. (2023). *Africa cyber capacity-building programme report*. Global Forum on Cyber Expertise.

Kazimierczak, M., et al. (2024). Impact of AI on the cyber kill chain: A systematic review. *Journal of Cybersecurity Research*.

Kenya National Cybersecurity Strategy. (2022). *Government of Kenya*.

Kephart, J., & Chess, D. (2003). The vision of autonomic computing. *IEEE Computer*, *36*(1), 41–50.

Kim & Cho (2023). *My teammate is AI: Understanding students' perceptions of student-AI collaboration in drawing tasks*. *Asia Pacific Journal of Education*, 1–15. https://doi.org/10.1080/02188791.2023.2286206

KPMG. (2023). *Cybersecurity readiness in Africa: Comparative country analysis*. KPMG International.

Krti, T. (2025). *Cyber Sentinel: An emergent threat detection system for AI security* [Preprint].

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

Mahmud, T., Prince, M. A. H., Ali, M. H., Hossain, M. S., & Andersson, K. (2024). Enhancing cybersecurity: Hybrid deep learning approaches to smishing attack detection. *Systems, 12*(11), 490.

Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and cognition: The realization of the living*. D. Reidel Publishing.

NIST (Vassilev, A., et al.). (2025). *Adversarial machine learning taxonomy (NIST AI Report)*. National Institute of Standards and Technology.

NIST. (2024). *AI Risk Management Framework (NIST AI 100-1)*. National Institute of Standards and Technology.

Oh, S. H., et al. (2024). Employing deep reinforcement learning for cyber-attack detection and response. *Electronics (MDPI)*.

Rahmati, M. (2025). *Towards explainable and lightweight AI for real-time cyber threat hunting in edge networks* [Preprint].

Reynaud, J. (2025). *The explainability–autonomy dilemma in AI governance*. *AI & Ethics*.

Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson Education.

Xu, M., et al. (2025). *LLM-based agents in autonomous cyberattacks*. *arXiv preprint*.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]