OPEN ACCESS

Kashf Journal of Multidisciplinary Research
Vol: 02 - Issue 12 (2025)
P-ISSN: 3007-1992 E-ISSN: 3007-200X
https://kjmr.com.pk

A SYSTEMATIC LITERATURE REVIEW OF AGENTIC AI SYSTEMS FOR AUTONOMOUS PREVENTION, DETECTION, AND MITIGATION OF CYBER ATTACKS IN CRITICAL INFRASTRUCTURE

***Irfan Ali*[1], *Zaheer Ahmed*[2], *Massna ijaz*[3], *Faisal Rehman*[4, 5], **Rizwan Hameed*[6], *Junaid Akbar*[4]**

[1, 2, 3]*University of Gujrat, Pakistan*
[4]*Department of Statistics & Data Science, University of Mianwali, Mianwali, Pakistan*
[5]*Department of Robotics & Artificial Intelligence, National University of Sciences and Technology (NUST), Islamabad, Pakistan*
[6]*Computer Science, School of Computing, Gold Campus, Superior University, Lahore, Pakistan.*

***Corresponding Author:** (rizwan.hameed98@gmail.com)*

***DOI:** (https://doi.org/10.71146/kjmr713)*

Article Info

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license https://creativecommons.org/licenses/by/4.0

Abstract

Critical infrastructure, including industrial control systems (ICS), agricultural IoT, smart grids and connected manufacturing, now faces attackers who combine AI tooling with zero-day exploits and operate at machine speed. Reactive, rule-based and human-supervised defences cannot keep pace with the volume and velocity of alerts. This paper reviews how agentic AI, meaning systems that plan, reason, use tools and act with limited human input, can defend such infrastructure autonomously. Following PRISMA 2020, we synthesise 25 studies spanning LLM-based defenders, reinforcement learning, edge-based federated learning, multi-agent systems, physical industrial testbeds and governance frameworks. We organise them in a taxonomy aligned with the NIST Cybersecurity Framework (CSF) 2.0 and compare them on detection accuracy, response latency, scalability and adversarial robustness. Reported results suggest that agentic and machine-learning defences outperform rule- and signature-based systems on the datasets used, with detection accuracy clustered at 93 to 99 percent and sub-second latencies common. These systems also introduce new risks, because attackers can target their reasoning, memory and tool use, and most results have yet to be shown to transfer from simulation to the field. We conclude that progress depends on hybrid neural-symbolic designs, on governance anchored in NIST CSF 2.0, and on validation in physical testbeds and field settings rather than in simulation alone. Recovery, agricultural IoT and shared physical-testbed benchmarks are the priority gaps, because failures there are the most costly.

Keywords: *Autonomous cyber defense, Agentic AI, Critical infrastructure, Industrial control systems, Smart agriculture, NIST Cybersecurity Framework, Federated learning, Reinforcement learning, Large language models, Threat mitigation.*

1 Introduction

Cyberattacks on critical infrastructure have become routine. Reports from 2024 and 2025 put the average organisation at more than 1,300 attacks per week, and intrusions into critical national infrastructure rose by roughly 140 percent in a single reporting period [1]. The consequences extend beyond financial loss: an attack on an irrigation network, a wastewater plant or a substation endangers public safety, food supply and social stability. Defenders are stretched thin. The global cybersecurity workforce is short by almost four million people [2], and Security Operations Centre (SOC) analysts spend most of their time triaging low-fidelity alerts that attackers can generate at almost no cost. Three assumptions behind the conventional defensive model no longer hold: that threats can be enumerated in advance, that humans can keep pace with alert volume, and that every response can wait for sequential approval [3]. Machine-speed attack tooling widens the gap further [3].

Agentic AI has emerged as a serious response, but the term needs care because it is often conflated with simpler ideas. An AI agent performs a single, bounded task, such as raising an alarm, summarising a packet capture or classifying a file as malicious. An agentic AI system does more: it pursues a stated objective, reasons through a sequence of steps, retains memory, invokes tools through structured interfaces and often coordinates with other agents [4]. This difference lets it go beyond flagging an anomaly: it can hypothesise a cause, isolate the affected component and report within seconds rather than days.

This review has three objectives. First, it surveys and consolidates existing work on agentic AI defences for critical infrastructure. Second, it organises that work in a taxonomy based on NIST CSF 2.0, so that practitioners can find evidence by function and domain. Third, it compares the reported performance of the reviewed approaches on detection accuracy, response speed, scalability and adversarial robustness. Two domains receive particular attention. ICS form the core of energy and manufacturing and combine legacy protocols with physical actuation [5, 6, 7, 8]. Agricultural IoT is often overlooked in the security literature yet is an increasingly exposed target as smart farming spreads; one survey of 1,800 farmers found that more than half had already been attacked [9]. Our central argument is that progress in both domains depends on hybrid systems that combine neural and symbolic reasoning, on governance built around proven frameworks, and on validation in real plants and fields rather than in simulators alone.

1.1 Research Questions

The review is organised around four research questions. **RQ1:** Which agentic architectures (LLM-based, RL-based, federated, multi-agent and hybrid) have been proposed to defend critical infrastructure, and how do they map to the NIST CSF 2.0 functions? **RQ2:** How do they compare on detection accuracy, response latency, scalability and adversarial robustness, and how strong is the validation behind the reported figures? **RQ3:** What new risks do agentic defenders introduce, and which governance mechanisms are proposed to bound them? **RQ4:** Where are the evidence gaps, particularly in ICS and agricultural IoT, and what should future work prioritise? Explicit answers are given in Section 8.

2 Background and Theoretical Foundations

2.1 AI Agents and Agentic AI

Operational definition. We define an agentic AI system as software that satisfies at least four of five properties (Hanif et al. [4]): (1) goal-directed autonomy, pursuing an open-ended objective rather than a single task; (2) persistent or shared memory that spans interactions; (3) self-reflective reasoning or planning before acting; (4) orchestrated use of tools across several interfaces; and (5) coordination with other agents through defined protocols. Systems meeting fewer than four, such as a single-task RL policy, a classifier or a non-reflective LLM responder, are treated as AI components rather than agentic systems, and where a study uses the word "agent" loosely we note the gap.

An AI agent observes an environment, decides and acts to achieve a goal. The idea is old, but current systems are built on large language models, reinforcement learning or both. A single agent typically handles one session with limited memory and a fixed toolset, and it reacts only to the input it receives. Agentic AI extends this in five directions [4]: greater autonomy, in which the agent sets sub-goals and reorganises tasks; persistent, shared memory; self-reflection, in which the agent critiques its plan before acting; parallel, orchestrated tool use; and multi-agent coordination through defined protocols. The result can pursue an objective such as "defend this network" through choices that no single prompt could specify. Defensive architectures fall into five families: RL policies trained in cyber-range simulators such as CybORG and CAGE [10, 11, 12, 13, 14]; LLM reasoning agents that read logs in natural language and choose responses [1, 10]; multi-agent systems in which a supervisor coordinates specialised sensing and analysis agents [15, 16]; federated learning agents that share model updates rather than raw data [1, 17]; and hybrids that combine deep learning, RL and symbolic logic [15, 18].

2.2 Cybersecurity Functions and the NIST Framework

NIST CSF 2.0 organises defence into six functions. Govern, new in version 2.0, makes strategy and policy explicit. Identify catalogues assets and vulnerabilities; Protect hardens systems with controls; Detect surfaces anomalies in real time; Respond contains and remediates incidents; and Recover restores normal operations and captures lessons learned [19]. Agentic defences fit Detect and Respond best. Recover is gaining coverage through self-healing and rollback mechanisms [20], Protect remains thinner because it usually requires system changes that an agent cannot make alone, and Identify and Govern are still largely beyond current agents. Metin et al. [19] extend CSF 2.0 to agentic risks, for example by treating model weights as intangible assets and tool revocation as a response capability.

2.3 Critical Infrastructure Domains

Critical infrastructure comprises the physical and digital systems whose destruction would harm public wellbeing: energy grids, water and wastewater, transport, hospitals, finance and food production. This review focuses on two domains. ICS sit between digital command and physical actuation. They comprise programmable logic controllers (PLCs), SCADA systems, distributed control systems and the protocols that connect them, including Modbus and DNP3. Air gaps once protected them, but Industry 4.0 connectivity, remote monitoring and cloud integration have ended that assumption [1]. Typical attacks are denial of service, false data injection, replay, man-in-the-middle and reconnaissance [6, 7]. A

distinctive constraint is that defensive action itself has physical consequences, so an over-aggressive automated response can do harm.

Smart agriculture, or agricultural IoT, combines sensors, drones, autonomous vehicles and farm-management software with cloud analytics [9, 16, 21]. Global food demand is projected to rise by 70 percent by 2050, which makes the sector strategic, yet security research has historically neglected it. Yield imagery, GPS data that guides autonomous machinery and remotely controlled irrigation schedules are all soft targets. The survey of 1,800 farmers by Campoverde Molina and Lujan Mora found that over half had already been victims of a cyberattack [9].

3 Research Methodology

The review follows the PRISMA 2020 guidelines and proceeded through four phases: identification, screening, eligibility assessment and inclusion. The protocol was not registered in PROSPERO, which we acknowledge as a limitation. The reporting below is detailed enough for another researcher to reproduce the search.

3.1 Search Strategy and Data Sources

On 15 April 2026 we searched five sources: IEEE Xplore (~145 records), Elsevier ScienceDirect (~210), ACM Digital Library (~95), SpringerLink (~120) and Google Scholar (~180 records, screened from the first 200 ranked results). Google Scholar served as a supplementary source because its results are not fully reproducible across users and dates; screening only the top 200 results follows common practice in software engineering reviews. The search covered 2021 to 2025, the period in which generative AI and LLMs became mainstream, and used the following Boolean string, adapted to each database's syntax:

> ("agentic AI" OR "autonomous agent" OR "LLM agent" OR "goal-oriented AI") AND ("critical infrastructure" OR "ICS" OR "industrial control system" OR "IoT" OR "agriculture") AND ("cybersecurity" OR "intrusion detection" OR "mitigation")

3.2 Inclusion and Exclusion Criteria

Inclusion criteria: (1) peer-reviewed journal or conference papers; (2) emphasis on autonomous or agentic decision-making; (3) application to critical infrastructure (energy, water, transport or agriculture); (4) full text available in English; (5) publication between January 2021 and December 2025. **Exclusion criteria:** (1) narrow AI that does not address autonomous goal setting; (2) non-English articles; (3) short abstracts or posters without full technical procedures; (4) duplicate findings from an author group already represented by a more complete study; (5) publications before 2021, except as foundational background.

3.3 Data Extraction

A structured template was applied to each of the 25 included studies. It recorded (i) bibliographic data; (ii) the technical problem addressed; (iii) the proposed agentic or AI-based solution; (iv) the architecture

family (LLM-based, RL-based, federated, hybrid or governance); (v) the experimental method, including datasets, testbeds and simulators; (vi) quantitative results (accuracy, precision, recall, F1, latency); and (vii) author-acknowledged limitations. The full extraction sheet is available in the supplementary protocol file.

3.4 Quality Assessment Rubric

Every study was assessed for its contribution to the field, with particular attention to the practicality gap, that is, the difference between theoretical models and those demonstrated on physical hardware or high-fidelity digital twins. Each study was scored on a ten-point rubric, awarding one point per item (half a point for partial satisfaction): (1) explicit research aims and questions; (2) alignment between study design and claims, particularly regarding real-world readiness; (3) transparency of the dataset or testbed, including accessibility and provenance; (4) an explicit baseline evaluated under identical conditions; (5) appropriate and adequate metrics, with accuracy-only reporting on imbalanced data penalised; (6) reproducibility, covering hyperparameters, seeds, code and environment; (7) validation realism on the simulation-to-reality spectrum, with physical testbeds scoring full marks and toy simulators zero; (8) adversarial or out-of-distribution evaluation rather than testing on the training distribution alone; (9) statistical reporting through variance, confidence intervals or multiple runs; and (10) explicit acknowledgement of limitations and threats to validity. Studies were banded as high (7 or more), medium (4 to 6.5) or low quality (3.5 or less). No study was excluded on quality alone, but higher-scoring studies carry more interpretive weight in Sections 4, 6 and 7.

3.5 Limitations and Sources of Bias

Five limitations apply. First, the protocol was not pre-registered, which limits external auditability. Second, a single reviewer carried out screening, extraction and quality scoring, which risks selection and interpretation bias; a second-reviewer check on a random subsample is the main strengthening step for a future revision. Third, Google Scholar is not fully reproducible, a risk only partly mitigated by screening the top 200 results. Fourth, English-only inclusion may have excluded relevant work from non-Anglophone communities. Fifth, the field is publishing quickly, and several included works are preprints or recent articles that may be revised. No meta-analytic pooling was performed because datasets, metrics and evaluation conditions are too heterogeneous (Section 7).

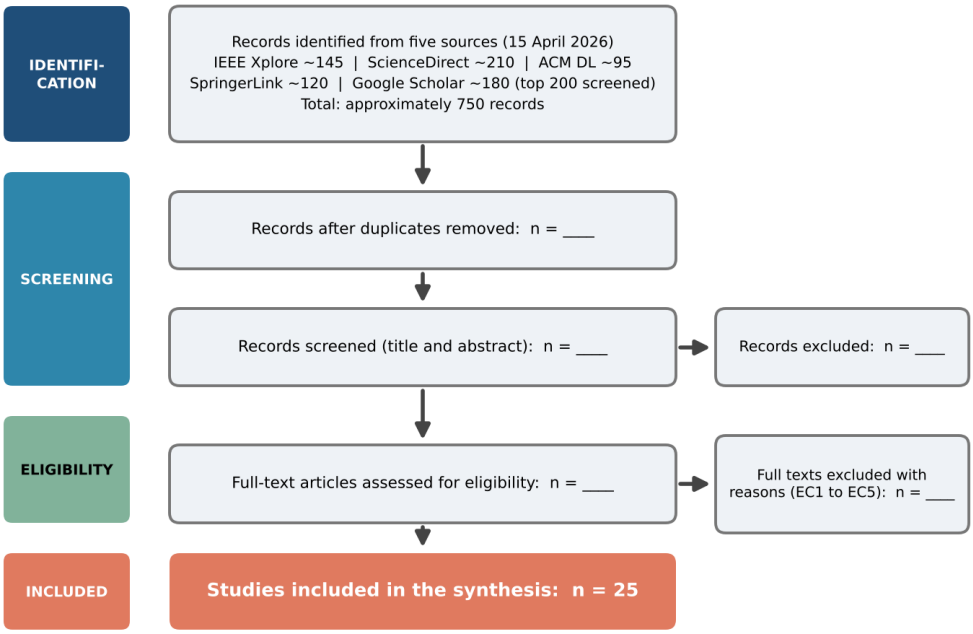


Fig. 1. PRISMA 2020 flow of study selection. Source counts and the final 25 included studies are as reported in Section 3; the intermediate stage counts are left to be completed from the original screening log.

4 Literature Review

Research on agentic AI for critical-infrastructure defence has grown rapidly. We examined 25 peer-reviewed studies, most published between 2021 and 2025, and grouped them into eight themes that mirror how the research community has approached the problem. Fig. 2 gives an overview of the review pipeline, the eight themes, their NIST coverage and the main findings. Each subsection ends with a critical assessment of the evidence.

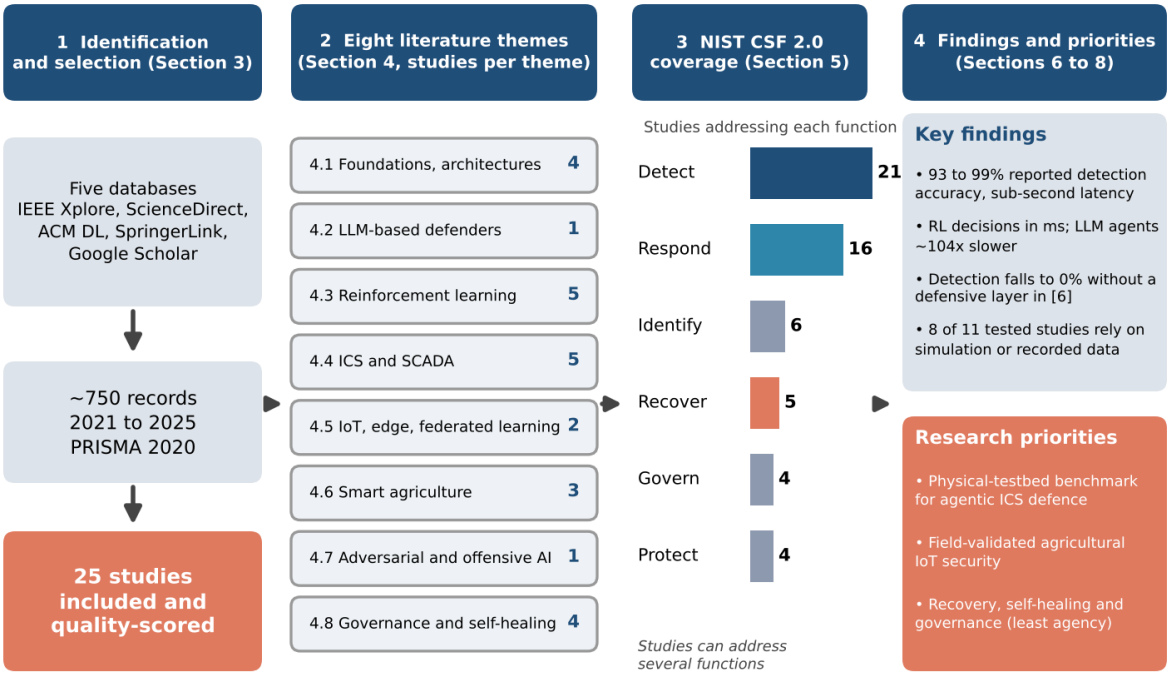


Fig. 2. Overview of the review. (1) Identification and selection (Section 3); (2) the eight literature themes with the number of studies in each (Section 4); (3) the number of studies addressing each NIST CSF 2.0 function (Section 5; one study can address several); (4) the main findings and research priorities (Sections 6 to 8).

4.1 Foundations and Architectural Models

Kshetri [2] offers one of the earliest broad framings of agentic AI in cybersecurity, drawing on commercial deployments at CrowdStrike, ReliaQuest and Darktrace. ReliaQuest GreyMatter is reported to process alerts twenty times faster than manual triage and to automate 98 percent of alert responses; Darktrace ENDPOINT cuts triage time by 98 percent; and Microsoft Security Copilot reduces mean time to respond by 30 percent. The market for AI in cybersecurity is projected to grow from 24.8 billion dollars in 2024 to 146.5 billion by 2034 [2]. Kott [5] proposes the Autonomous Intelligent Cyber defense Agent (AICA) reference architecture, now a benchmark for the sub-domain. It defines five functions (sensing and world-state identification, planning and action selection, collaboration with friendly agents, action execution, and learning) and stresses adversarial reasoning and cyber camouflage: the agent itself must be hard for an attacker to detect. Hanif et al. [4] formalise the gap between AI agents and agentic AI through five pillars, and Adabara et al. [18] extend this into a multi-dimensional framework that adds quantum-resilient defence and ethical governance. Together these works supply the vocabulary used across the rest of the literature.

4.2 LLM-Based Autonomous Defenders

Sebastian et al. [10] integrate LLMs into the CybORG CAGE 4 environment through an extensible adapter that converts observation vectors into natural language, testing GPT-4o mini, o1-mini, o3-mini and DeepSeek V3. o3-mini earns the highest reward among the LLM agents, yet an RL-only team still outperforms every LLM team (mean reward -493 versus -2,547), and RL agents select actions about 104

times faster than the fastest LLM. The paper also introduces an eight-bit protocol for inter-agent messages, a small but useful contribution to multi-agent autonomous cyber defence.

Critical assessment. The evidence here rests mainly on one simulation study (CybORG CAGE 4) and a small set of general-purpose LLMs [10]. The finding that RL beats LLM teams roughly fivefold on reward [10] is striking but tied to CybORG's discrete action space; whether it generalises is untested. LLM-based defenders in the corpus are otherwise evaluated mainly in simulation [6], so their behaviour on real industrial protocols remains an open question.

4.3 Reinforcement Learning for Autonomous Cyber Defence

Wolk et al. [11] evaluate PPO variants in CAGE Challenge 2, including hierarchical RL, action masking, custom training and ensembles. Their multi-ensemble agent reaches a total reward of -57.14, against -95.47 for the PPO baseline, -823.89 for a heuristic defender and -2,015.61 for random defence, with an average decision time of 6.67 ms, fast enough for real-time use. A counter-intuitive result follows: defenders trained against efficient attackers do worse against a less efficient but more random one, because randomness is harder to predict than competence. Collyer et al. [12] address the simulation-to-reality gap by adding FEATHER-G graph embeddings of the network to the agent's observation space. In YAWNING TITAN, graph embedding outperforms a flattened adjacency matrix in unseen 20-node environments and reaches the maximum reward (599) in 40-node environments where the standard agent peaks at 430; the standard representation remains better in simple 10-node settings, so embeddings pay off where topology varies. Prebot et al. [13] apply Instance-Based Learning Theory to motivate AICA decision-making and compare 40 simulated IBL agents with 120 human defenders in an Interactive Defence Game. Both humans and agents suffer more from a directed Beeline attacker than from a stochastic Meander attacker, and both show clear learning curves, a reminder that defender behaviour is itself something to model. Tran and Akella [14] tackle scale with HA-DRL, a hierarchical agent that decomposes a 4,646-action space over 100 hosts; a standard Double DQN learns nothing there, whereas HA-DRL reaches an optimal policy in about 4,000 episodes. Rajgopal [8] applies DQN with policy-gradient enhancements to a custom simulation built on DARPA and CICIDS2017 data and reports 95 percent detection with mitigation latency below 50 ms in federated settings.

Critical assessment. Rewards across CAGE-family studies are not directly comparable because the environment changed between editions. The generalisation results [11, 12] deserve particular weight: nearly every study in this strand tests against a few predefined attacker policies, and the performance drop against unpredictable attackers [11] suggests the field is over-optimised for learnable opponents.

4.4 Industrial Control Systems and SCADA

Shrestha [6] provides one of the most rigorous empirical studies of agentic AI in industrial automation. A PRISMA-driven review of 46 articles is followed by a proof of concept in which CrewAI and LangFlow face four attack types: DDoS, false data injection, replay and adversarial attacks. Without a GAN defence, detection falls to zero and the false-negative rate rises to 1.0. With a GAN discriminator and GPT-4o mini, CrewAI reaches 67 percent accuracy (precision 1.00, recall 0.50, AUC-ROC 0.95) and LangFlow trails at 60 percent, giving a clean comparison of how framework architecture shapes

attack resilience. Houkan et al. [7] take a hardware-first approach: a physical testbed of Mitsubishi R04ENCPU PLCs and SCALANCE switches, seven attack types, and ICSCASD-MPLC, a labelled dataset of 2.6 million records with 57 features. XGBoost reaches 99 percent binary and 97.3 percent multi-class accuracy; a Decision Tree reaches 97 and 98.6 percent. To our knowledge this dataset is the most solid empirical basis for ICS-oriented machine learning in the corpus. Mitchell et al. [15] propose a four-layer agentic model (perception, reasoning, decision-making, actuation) for national critical infrastructure and report detection above 90 percent with fewer false positives through hybrid reasoning. Gvilava and Tsinadze [22] examine decision-support systems across architectural levels and stress resilience under incomplete information. Bondhala [23] reports 95 percent detection for neural-network IDS against 57 percent for traditional systems, and AI automation that cuts mean time to detect by 74 percent and mean time to respond by 64 percent.

Critical assessment. An accuracy of 99 percent on a single testbed [7] marks a clear empirical advance but should be read with care: it usually reflects performance against the attack scripts used during data collection and may not generalise to adversarially crafted attacks. The "above 90 percent" detection reported by Mitchell et al. [15] lacks dataset size, class balance and baseline details, so it cannot be positioned against the Houkan results [7].

4.5 IoT, Edge, and Federated Learning

Toree et al. [17] address privacy in federated IDS. Their 1D CNN averages 97.31 percent across seven IoT datasets (including TON_IoT, BoT-IoT and CIC IoMT 2024, over 415 million records) while combining Laplace-Gamma differential privacy, Diffie-Hellman key exchange and fully homomorphic encryption at a reported computational overhead of about 10 percent, the kind of trade-off that makes real deployment plausible. Ajayi and Masunda [24] report edge-AI deployments in smart manufacturing (detection under 500 ms, 85 percent breach reduction), healthcare (under 1 second, 72 percent) and military operations (under 300 ms, 90 percent). For non-agentic context, a systematic review of ML- and DL-based intrusion detection for IoT [25] and a lightweight one-class approach to IoT botnet detection [26] describe conventional IoT intrusion-detection approaches against which the federated and edge designs above can be read.

Critical assessment. The 10 percent overhead claimed for layered differential privacy, key exchange and FHE [17] is significant if independently reproduced, but it is averaged across seven datasets of very different size and feature space, so per-dataset overhead probably varies more than the mean suggests. The edge-AI figures of Ajayi and Masunda [24] come from case studies rather than controlled experiments, so they indicate the order of magnitude of what is achievable but cannot be compared directly with controlled experiments.

4.6 Smart Agriculture

Murad and Ahmed [16] built one of the most tangible agentic products in agriculture: four agents handle soil monitoring, weather, vision-based crop-disease detection and a supervisory chat interface in regional languages, and a custom robot actuates in the field. On 12,000 sensor samples from fields in Karachi and Thatta, Pakistan, the soil 1D-CNN reaches 96 percent accuracy, the weather model achieves

a mean absolute error of 0.27 for temperature, and the MobileViT vision model attains 98.5 percent training and 93.5 percent cross-validation accuracy for rice disease. The work is grounded in field conditions in a way most agricultural AI papers are not. Alahe et al. [21] and Campoverde Molina and Lujan Mora [9] both offer systematic reviews of smart-agriculture cyber threats. Alahe et al. classify threats by network layer and identify GPS data integrity and real-time processing as critical gaps; Campoverde Molina and Lujan Mora find that more than half of the relevant research appeared in the last two years and report an IDS reaching 99.92 percent accuracy for binary attack classification in agricultural IoT. Both argue for adopting NIST CSF 2.0 in agriculture, which is still rare. A recent chapter on AI integration in drone technology spans agricultural and security applications [27], a reminder that the aerial and sensing platforms used in precision farming are also part of its attack surface.

Critical assessment. The agricultural strand is under-populated relative to its strategic importance. Murad and Ahmed [16] is one of the few field-grounded agentic systems in the whole corpus, but its soil and weather models are single-region (Karachi and Thatta), its rice-disease model relies on a Kaggle dataset with unverified label provenance, and the five-point gap between training (98.5%) and cross-validation (93.5%) accuracy is consistent with mild overfitting that would likely widen in cross-region deployment.

4.7 Adversarial and Offensive Agentic AI

Karamchand and Aramide [3] place AI cyberwar in strategic context, charting state investment (United States 120 billion dollars, China 110 billion, EU 90 billion, Russia 70 billion) and describing AI as a third revolution in military affairs comparable to nuclear weapons. Muneer et al. [28] discuss security and privacy concerns in AI models more broadly, and an examination of phishing attempts on web-based business applications [29] gives conventional background on phishing attacks.

Critical assessment. Evidence in this strand is largely strategic or conceptual. The corpus contains few direct experiments on attacks against agentic defenders: the closest are the adversarial tests of CrewAI and LangFlow in [6] and the agentic threat taxonomy of [19]. Claims about how vulnerable agentic systems are in deployment should therefore be read as open questions, not as measured rates.

4.8 Governance, Verification, and Self-Healing

Joshi [30] documents the policy gap: frameworks addressing agentic-AI compliance risk exist in only 23 percent of organisations, even though 67 percent of Fortune 500 enterprises were expected to deploy agentic AI by the end of 2025. The study proposes the RSI-Optimize algorithm, reported to reduce safety violations to 0.2 percent against 3.7 percent for a baseline. Metin et al. [19] apply NIST CSF 2.0 to agentic AI and propose a four-dimensional taxonomy covering reasoning and planning, memory and state, actions and tool use, and multi-agent interaction; they promote a Principle of Least Agency, analogous to least privilege. Yigit and Ferrag [1] propose a five-step LLM lifecycle for critical infrastructure protection (vision, selection, performance adjustment, evaluation, deployment) and a matrix contrasting adaptive thresholds and automated isolation in agentic protection with fixed thresholds and manual inspection in traditional protection. Ibitoye [20] focuses on self-healing networks,

reporting 80 percent of legitimate traffic preserved under DDoS, 65 percent shorter recovery time in a financial-trading simulation and a 70 percent smaller zero-day exposure window through automated patching.

Critical assessment. Governance work in the corpus is descriptive and normative rather than empirical. The RSI-Optimize result (0.2% against 3.7%) [30] comes without enough detail on the evaluation harness to tell whether it is a stable result or a single-run estimate, and we treat it accordingly in Sections 5 and 7.

5 Research Analysis and Taxonomy

We classify the reviewed work along two axes in common practitioner use: the NIST CSF 2.0 function addressed (Identify, Protect, Detect, Respond, Recover, Govern) and the application domain (ICS, IoT and edge, agricultural IoT, network and cloud, or general critical infrastructure). Table 1 classifies all 25 studies and labels each with its main technical method and validation type, because the simulation-to-reality gap is among the most important variables in the field.

Table 1. Comparative classification of agentic AI cybersecurity work across NIST functions, domain, approach, and validation.

Ref	Year	Domain	NIST Function(s)	Approach	Validation	Key Result
[1]	2025	General CI	Detect, Respond, Recover	LLM lifecycle for CIP	Meta analysis	5 step CIP lifecycle, benchmark map
[2]	2025	General CI	Detect, Respond	Agentic AI survey, RMF	Industry data	98% alert automation, 30% MTTR reduction
[3]	2025	Cyberwarfare	All functions	Strategic analysis	Synthesis	AI as third revolution in military affairs
[4]	2025	General AI	N/A (foundations)	Comparative conceptual evaluation	Architectural analysis	5 pillar agentic AI taxonomy
[5]	2023	General CI	Identify, Protect, Detect, Respond, Recover	AICA reference architecture	Concept + prototype	5 function reference architecture
[6]	2025	ICS	Detect	GAN + LLM, CrewAI/LangFlow	Simulation	67% acc with GAN, 0% without
[7]	2025	ICS	Detect	DT, XGBoost on physical PLCs	Mitsubishi PLC testbed	99% binary, 2.6M record dataset
[8]	2025	General CI	Detect, Respond	DQN + policy gradient + LLM	DARPA, CICIDS2017	95% acc, <50 ms federated latency
[9]	2025	Smart agriculture	Identify, Protect	Systematic review	PRISMA, 58 documents	50% farmers attacked, 99.92% IDS acc
[10]	2025	General CI	Detect, Respond	LLM agents in CybORG CAGE 4	Simulation	RL outperforms LLM by 5x reward
[11]	2022	Network defence	Detect, Respond	PPO ensemble RL	CAGE Challenge 2	Reward -57.14, 6.67 ms decision
[12]	2022	Network defence	Detect	Graph embedded RL (FEATHER G)	YAWNING TITAN	Generalises in 20 and 40 node networks
[13]	2022	Network defence	Detect, Respond	Instance Based Learning + IDG	120 humans + 40 agents	Human aligned learning curves
[14]	2021	Pen testing	Identify	Hierarchical DRL (HA DRL)	Simulation	100 host scaling, DDQN fails baseline
[15]	2025	National CI	Detect, Respond	Hybrid DL + RL + symbolic	High fidelity testbed	>90% accuracy, hybrid reasoning
[16]	2025	Smart agriculture	Detect, Respond	Multi agent + custom robot	Field data Pakistan	98.5% disease detection MobileViT
[17]	2025	IoT	Detect	FL + 1D CNN + DP + FHE	7 datasets, 415M records	97.31% accuracy, 10% overhead

Ref	Year	Domain	NIST Function(s)	Approach	Validation	Key Result
[18]	2025	General CI	Detect, Respond, Govern	Cognitive autonomy + quantum resilience	Narrative review	11 category technological taxonomy
[19]	2025	Agentic AI security	Govern, Identify, Protect, Detect, Respond, Recover	NIST CSF 2.0 adaptation	PRISMA SLR, 30 studies	4 domain agentic threat taxonomy
[20]	2021	Network	Detect, Respond, Recover	Self healing AI architecture	Conceptual + cases	65% recovery time reduction
[21]	2024	Smart agriculture	Identify, Detect	Layered threat framework	PRISMA review	Layer wise threat classification
[22]	2025	Energy, transport, health	Detect, Respond	Multi layer autonomous DSS	Architectural model	Layered fault tolerance design
[23]	2024	Data centres	Detect, Respond	Deep learning IDPS, Zero Trust	Industry data	95% detection, 74% MTTD reduction
[24]	2025	Manufacturing, health, military	Detect, Respond	Edge AI + federated learning	Case studies	<500 ms detection, 85% breach reduction
[30]	2025	Policy	Govern	Three pillar governance model	SLR + math modelling	0.2% safety violations w/ RSI Optimize

Note on cross-study comparability. The studies in Table 1 differ in datasets, attack taxonomies, hardware and evaluation protocols. Reported figures therefore describe each study individually and should not be subtracted, ranked or pooled. Summary statements about trends (for example, that accuracies cluster in the 93 to 99 percent range) describe the distribution of reported figures, not a normalised comparison, and no meta-analytic adjustment for dataset difficulty, class imbalance or evaluation realism has been made.

5.1 Trends, Strengths, Weaknesses, and Gaps

Table 1 shows several clear trends. Detect dominates: 21 of the 25 studies list it, with Respond a frequent secondary focus. Recover is almost uncovered and is the primary subject of a single study [20], while Identify and Govern remain the weakest functions [19, 30]. Methodologically the literature falls into three strands: a large simulation strand built on CybORG, CAGE and YAWNING TITAN [8, 10, 11, 12, 13, 14] that gives clean results but limited transferability; a smaller strand using public IoT datasets [17]; and a still smaller physical-testbed strand [7, 15] that is highly valuable but costly to replicate. In our corpus only three of the 25 studies rest on a physical testbed or field deployment [7, 15, 16].

Agentic systems deliver consistent gains over rule- and signature-based defences. Detection accuracy often falls in the 93 to 99 percent range on heterogeneous data, and most designs respond in well under a second. Federated learning is now practical: the combination of differential privacy, key exchange and homomorphic encryption in [17] costs only about 10 percent extra compute. Combining neural pattern recognition with symbolic logic [15] gives hybrid architectures that cut false positives without sacrificing detection. The weaknesses are equally consistent. Without a defensive layer, performance under attack collapses: in [6] both CrewAI and LangFlow fall to 0 percent accuracy without the GAN defence. LLM agents are slow compared with RL policies [10]. Generalisation to new topologies and unpredictable attackers remains poor [11, 12].

Three gaps stand out. First, no agreed benchmark exists for agentic ICS defence using real industrial protocols and physical PLCs; Houkan et al. [7] move in that direction, but the community lacks a

common evaluation surface. Second, agricultural IoT is repeatedly described as a soft target [9, 16, 21], yet few studies go beyond general-purpose IDS. Third, recovery and self-healing lag far behind detection: most agentic systems can detect an attack but cannot return a partly damaged plant to a safe operating state without human help.

5.2 Evidence-Strength Map

Table 5 counts the studies in Table 1 by NIST function and by the type of validation behind them, which shows where the evidence is firm and where it is thin. Physical or field testbeds are the strongest evidence, followed by labelled datasets and then simulation; reviews, frameworks and case studies are grouped as conceptual evidence. The grouping of validation types is the reviewers' own and a study is counted under every function it lists.

Table 5. Number of studies by NIST CSF 2.0 function and validation type, derived from Table 1.

NIST function	Studies	Physical / field testbed	Dataset-based	Simulation / experiment	Review / conceptual / other	Evidence strength
Govern	4	0	0	0	4	Normative
Identify	6	0	0	1	5	Weak
Protect	4	0	0	0	4	Weak
Detect	21	3	2	5	11	Strongest
Respond	16	2	1	3	10	Limited
Recover	5	0	0	0	5	Very thin

Physical-testbed or field evidence is concentrated in Detect (three studies) and is almost absent elsewhere (two studies for Respond, none for the other four functions), so Respond and Recover, the functions where a wrong autonomous action costs most, rest mainly on simulation, datasets or conceptual work.

6 Discussion

Across the 25 studies, critical-infrastructure defence is shifting from reactive, signature-based detection towards proactive, closed-loop defence built on autonomous reasoning and tool use. Four themes emerge.

6.1 The Performance Trade-off: LLMs versus Reinforcement Learning

LLM and RL agents play distinct roles. Sebastian et al. [10] and Yigit and Ferrag [1] show that LLM agents excel at high-level reasoning, protocol analysis and explaining complex attack paths, but their latency limits use in real-time ICS. Deep RL agents, as in the CAGE studies [11, 12], act in milliseconds yet lack the semantic understanding to handle unfamiliar, non-tabular threats. The evidence points to hybrid architectures, in which LLMs orchestrate strategy and RL executes fast mitigation, as the most resilient design for critical infrastructure.

6.2 Domain-Specific Resilience: Agriculture and ICS

Smart agriculture is a key frontier. Murad and Ahmed [16] show that decentralised agentic systems can cope with weak rural connectivity: edge-resident agents, such as those on the field robot, keep farm

systems secure and operational when cloud servers are unreachable. In the wider ICS context, autonomous decision-support systems [22] are reported to cut operator cognitive load by 30 to 40 percent, particularly in multi-vector attacks where manual triage is no longer feasible.

6.3 The Simulation-to-Reality Gap

Reported scores across the corpus are high, commonly above 95 percent on detection tasks, yet a critical practicality gap remains. Of the eleven reviewed studies that report an experimental evaluation, eight (about 73 percent) validate their systems in simulated environments or on pre-recorded datasets rather than on physical infrastructure [6, 8, 10, 11, 12, 13, 14, 17]. The same applies to this review: our conclusions on latency, scalability and accuracy rest on a corpus in which physical-testbed evidence is the minority [7, 15, 16]. We therefore treat the strongest empirical claims, particularly end-to-end response latency and adversarial robustness in production, as upper bounds from controlled environments, not as established field performance. The studies grounded in physical hardware or a high-fidelity testbed [7, 15] and the field-deployed agricultural framework [16] are weighted more heavily for this reason, and the call for shared physical-testbed benchmarks in Section 8 follows directly from this limitation.

6.4 Adversarial Risks and Governance

The review surfaces an autonomy paradox: increasing the capability of defensive agents also increases their value as targets, because an agent's reasoning logic becomes a high-leverage attack surface in its own right [6, 19]. This creates an apparent tension between literature that advocates autonomous, closed-loop defence and literature that insists on human oversight, particularly in industrial settings where defensive actions have physical consequences [6, 15, 19]. We argue that the two positions are not contradictory but operate at different layers (Fig. 3). Autonomy is appropriate at the detection and time-critical containment layer, where machine-speed response is operationally necessary and a delayed human decision is itself a failure mode. Human oversight is appropriate at the policy, escalation and recovery layer, where physical consequences, accountability and contextual judgement dominate. The Principle of Least Agency of Metin et al. [19] formalises this layering: each agent receives the minimum scope of autonomy its function requires, with hard policy bounds and human-in-the-loop overrides for any action that reaches into physical actuation or system-wide reconfiguration. The position of this review is therefore not autonomous defence versus human oversight, but autonomous defence within bounded, governed agency. Interpretability supports this oversight: explainable multi-scale feature fusion, developed for tumour detection, exposes which features drive a decision [31], and the same principle lets a human operator audit an automated defensive action before it is approved.

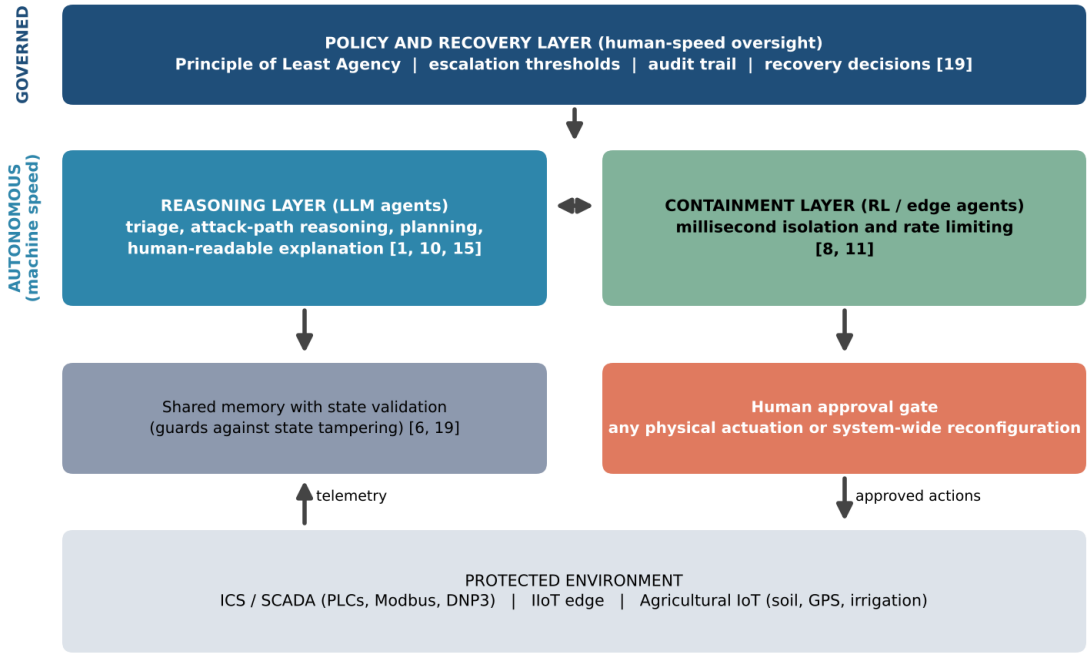


Fig. 3. Layered architecture for autonomous defence within bounded agency, synthesising Sections 6.1 and 6.4. LLM agents reason and explain, RL and edge agents contain at machine speed, and a governance layer with a human approval gate bounds any action that touches physical processes.

6.5 Threats to Validity

Table 6 groups the main threats to the validity of this review by type and states how each is addressed or disclosed.

Table 6. Threats to the validity of this review and how they are addressed.

Type	Threat to validity	Mitigation / status
Internal	Single-reviewer screening, extraction and quality scoring	Disclosed (Section 3.5); a second-reviewer check on a random subsample is the main planned strengthening step
Internal	Protocol not pre-registered	Disclosed; registration on OSF or PROSPERO is recommended for a future version
Construct	The term "agentic" is used loosely in the literature	Operational definition in Section 2.1 (at least four of five properties)
Construct	Accuracy hides recall on imbalanced data	Accuracy-only reporting is penalised in the rubric; recall and false-negative behaviour discussed in Section 7.1
External	Corpus dominated by simulation and recorded datasets	Physical-testbed studies weighted more; figures treated as upper bounds (Section 6.3, Table 5)
External	Metrics are not comparable across studies	Descriptive comparison only; no pooling or ranking (Section 5)
Coverage	English-only inclusion, non-reproducible Google Scholar, fast-moving field	Disclosed; top-200 screening; complementary context references in Section 4

6.6 Practitioner Guidance

Table 7 turns the findings into guidance for the main audiences of agentic defence.

Table 7. Practical guidance derived from the reviewed evidence.

Audience	Recommended practice	Evidence base
ICS / OT operators	Let agents detect and recommend; require human approval or hard policy bounds for any actuation or isolation of a physical process	[6], [15], [19]
SOC teams	Use RL or lightweight models for time-critical containment and LLMs for triage, explanation and planning; give each agent least privilege	[8], [10], [11]
Smart-farm operators and agri-tech vendors	Prefer edge-resident detection that keeps working offline; secure GPS and irrigation control paths first; adopt NIST CSF 2.0 profiles	[9], [16], [21]
Security architects	Treat the agent itself as an attack surface: validate memory and state, limit each agent's tools and autonomy (least agency) and test against adversarial inputs before deployment	[6], [19]
Regulators and standards bodies	Define escalation thresholds and audit-trail requirements for autonomous response	[19], [30]

7 Evaluation

We compare the reviewed approaches on four metrics that matter to practitioners: detection accuracy, response latency, scalability (to topology and load) and adversarial robustness. We then review domain-specific outcomes in ICS and agricultural IoT and the constraints that limit deployment today.

7.1 Detection Accuracy Across Approaches

Table 2. Figures are reproduced as reported in each cited study; cross-study comparison is descriptive only and is not statistically normalised.

Approach	Reference	Domain / Dataset	Accuracy / F1	Notes
CrewAI + GAN + GPT 4o mini	[6]	Industrial automation sim	67% (1.00 prec, 0.50 rec)	AUC ROC 0.95
LangFlow + GAN + GPT 4o mini	[6]	Industrial automation sim	60% (1.00 prec, 0.43 rec)	AUC ROC 0.88
Decision Tree	[7]	ICSCASD MPLC (PLC)	97% / 98.6%	Binary / multi class
XGBoost	[7]	ICSCASD MPLC (PLC)	99% / 97.3%	Binary / multi class
Rule based / signature	[8]	DARPA, CICIDS2017	82%	Traditional baseline
DQN + policy gradient + LLM	[8]	DARPA, CICIDS2017	95%	F1 93.5%
Hybrid DL + RL + symbolic	[15]	National CI testbed	>90%	Hybrid reasoning
MobileViT (rice disease)	[16]	Field data Pakistan	93.5% (CV) / 98.5% (train)	Smart agriculture
FL 1D CNN + DP + FHE	[17]	7 IoT datasets, 415M records	97.31% avg, 100% BoT IoT	Privacy preserving
Conventional rule based IDS	[23]	Data centre traffic	57%	Baseline for comparison
Deep learning IDPS	[23]	Data centre	95%	Behavioural analytics

Table 2 gathers detection results from studies that test against attack data with explicit ground-truth labels. Traditional rule- and signature-based systems converge at 57 to 82 percent. Machine-learning classifiers trained on labelled PLC traffic reach 97 to 99 percent [7]. With suitable defensive layers, modern agentic and federated methods push the upper bound above 97 percent on heterogeneous data. Two observations follow, both subject to the comparability caveat. First, within a single study the gap between strong and weak detection depends more on the defensive layering around the model than on the model family: in [6] the same agentic framework moves between 0 and 67 percent detection on identical attacks depending on whether a GAN discriminator is loaded. Because this is a within-study comparison, it is methodologically sound. Second, headline accuracy hides recall. Several agentic systems with strong overall accuracy still miss about half of attacks under adversarial conditions [6], and where a missed attack has physical rather than informational cost, recall and false-negative rate matter more than accuracy.

7.2 Response Latency and Containment

Table 3 summarises latency. RL policies are extremely fast: 6.67 ms for ensemble PPO [11] and under 50 ms in federated DQN setups [8]. LLM agents are orders of magnitude slower: Sebastian et al. [10] found the fastest LLM about 104 times slower than the fastest RL agent. Operational deployments therefore increasingly assign time-critical actions to RL and leave explanation, summarisation and high-level planning to LLMs. Work on real-time recognition with hybrid attention and transformer-enhanced feature fusion [32] suggests that richer models can be kept compact enough for fast inference, so part of the LLM latency gap may narrow as such designs mature.

Table 3. Decision latency and containment time across the reviewed agentic systems.

System / Method	Reference	Decision Latency	Containment Time	Recovery Notes
ReliaQuest GreyMatter	[2]	20x faster than manual	<5 minutes	98% alert automation
Microsoft Security Copilot	[2]	n/a	30% MTTR drop	Commercial LLM agent
DQN federated	[8]	<50 ms	Sub second	Mitigation latency
Fastest LLM (GPT 4o mini)	[10]	~700 ms (104x RL)	Multi step	Slow but interpretable
Ensemble PPO (CAGE 2)	[11]	6.67 ms	Within episode	RL only, fastest
Self healing networks	[20]	n/a	Adaptive isolation	65% recovery time reduction
AI driven SOC (data centre)	[23]	n/a	MTTR -64%	MTTD -74%
Edge AI (manufacturing)	[24]	<500 ms	n/a	85% breach reduction
Edge AI (military)	[24]	<300 ms	n/a	90% breach reduction
RSI Optimize (governance)	[30]	820 ms	n/a	0.2% safety violations

Containment times are reported less consistently. Federated DQN agents [8] report sub-second mitigation, a commercial SOC platform reports containment in under five minutes [2], and edge-AI deployments report detection in under 500 ms in manufacturing and under 300 ms in military settings [24]. Because some studies report decision time, some detection time and some end-to-end containment, these figures describe each system on its own terms and cannot be ranked (Section 7.6).

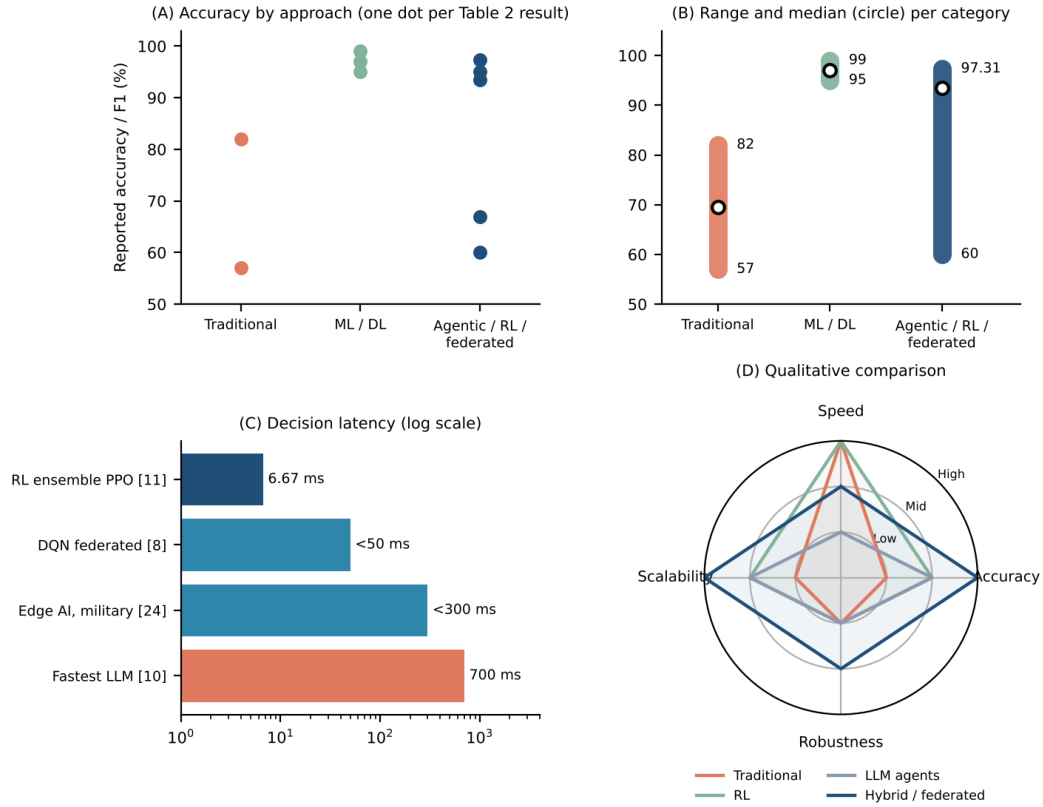


Fig. 4. (A) Reported accuracy by approach, one dot per Table 2 result (the >90% hybrid result [15] is a lower bound and is not plotted). (B) Range and median of those results per category; the two lowest agentic values are the CrewAI (67%) and LangFlow (60%) results of [6]. (C) Decision latency on a log scale: RL is fastest and LLM agents slowest. (D) Qualitative comparison of four approaches on accuracy, speed, scalability and robustness (the reviewers' summary of Sections 7.1 to 7.4, not a measured score).

7.3 Scalability and Adaptability

Scalability exposes the training-to-reality gap through topology. Wolk et al. [11] found that every PPO variant performed worse in unseen environments than in its training topology, and that host-configuration changes caused far more trouble than changes in attack paths, so the gap between simulation and deployment grows with environmental variation. Collyer et al. [12] showed that graph embeddings help in medium-sized networks (20 to 40 nodes) and slightly hurt in simple ones. Tran and Akella [14] showed that hierarchical RL becomes essential at scale: a plain DDQN cannot learn in a 100-host, 4,646-action environment, while HA-DRL learns in about 4,000 episodes. Adaptability, the ability to handle unseen attackers, is weaker: defenders trained against efficient attackers (Beeline and Meander) did even worse against a random attacker such as RandomMeander [11], because unpredictability undermines models that rely on learned attack patterns. The lesson is that evaluation must use unseen topologies and unseen attackers, because testing on the training distribution overstates performance.

7.4 Adversarial Robustness

Table 4 summarises adversarial findings, and the evidence is thin. In the one experimental study of attacks on agentic industrial agents in the corpus [6], detection falls to zero without a GAN defence and

reaches only 67 percent (CrewAI) and 60 percent (LangFlow) with one, so even the defended systems miss a large share of attacks. Bondhala [23] reports a 243-day breach detection gap for polymorphic malware in data centres, and 95 percent detection for deep-learning behavioural analytics against 57 percent for conventional rule-based systems. The agentic threat taxonomy of Metin et al. [19] and the decision-drift risk identified in [6] show where further attacks are likely, but the corpus offers no direct measurements of prompt injection, memory poisoning or goal hijacking against defensive agents. That is a gap in the evidence, not evidence that the risk is small.

Table 4. Adversarial vulnerabilities and mitigation strategies reported in the literature.

Vulnerability / Threat	Reference	Reported Rate / Outcome	Mitigation Tested
False data injection (ICS)	[6]	Detection 0% without GAN	GAN discriminator, LLM cross check
Decision drift in autonomy	[6]	Unsafe physical decisions	STRIDE based threat modelling
Polymorphic malware (data centre)	[23]	243 day breach detection gap	AI behavioural analytics

7.5 Domain-Specific Findings: ICS and Agricultural IoT

ICS have the strongest empirical foundation in the corpus. The Houkan dataset [7] holds 2.6 million labelled records from a physical PLC testbed, and 99 percent binary accuracy on it suggests that detection of known attack types is largely solvable offline. The harder problem is action: automated containment that isolates a faulty component may do as much damage as the attack itself. Mitchell et al. [15] address this by adding a symbolic-logic layer that explains each decision in human-readable form before it is executed. Agricultural IoT is less mature. The Murad and Ahmed framework [16] is one of the few works spanning concept to field deployment, with 96 percent accuracy on soil monitoring, a mean absolute error of 0.27 on temperature and 93.5 percent cross-validation accuracy on rice disease. Campoverde Molina and Lujan Mora [9] report 99.92 percent IDS accuracy for binary attack classification in agricultural networks, yet note that the field only began publishing in volume in 2023 and 2024. The distance between what is technically available and what is deployed is wider here than in any other critical-infrastructure area we reviewed.

7.6 Limitations and Possible Improvements

Most studies evaluate against attack patterns seen in training even while claiming open-world conditions; the generalisation tests in [11, 12] show how far performance falls once topologies and attackers change. Cross-study comparison is hampered by inconsistent definitions, especially of latency, where some authors report decision time and others end-to-end containment, and no agreed agentic ICS benchmark exists. The literature itself proposes three improvements. First, hybrid neural-symbolic architectures reduce false positives without losing accuracy, the right trade-off where false alarms are costly [15, 24]. Second, federated learning at the edge reduces privacy exposure and bandwidth use at a feasible overhead [17, 24]. Third, governance based on NIST CSF 2.0 [19] and sector standards (for example ISO/IEC in agriculture [9]) supplies the framework that agentic deployments lack.

7.7 Statistical Analysis

Fig. 5 summarises the corpus: most studies (17 of 25) date from 2025, general critical-infrastructure studies form the largest domain group, Recover receives minimal attention, and physical-testbed or field

evidence exists only for Detect and Respond, so the highest reported accuracies come mainly from offline datasets and single testbeds.

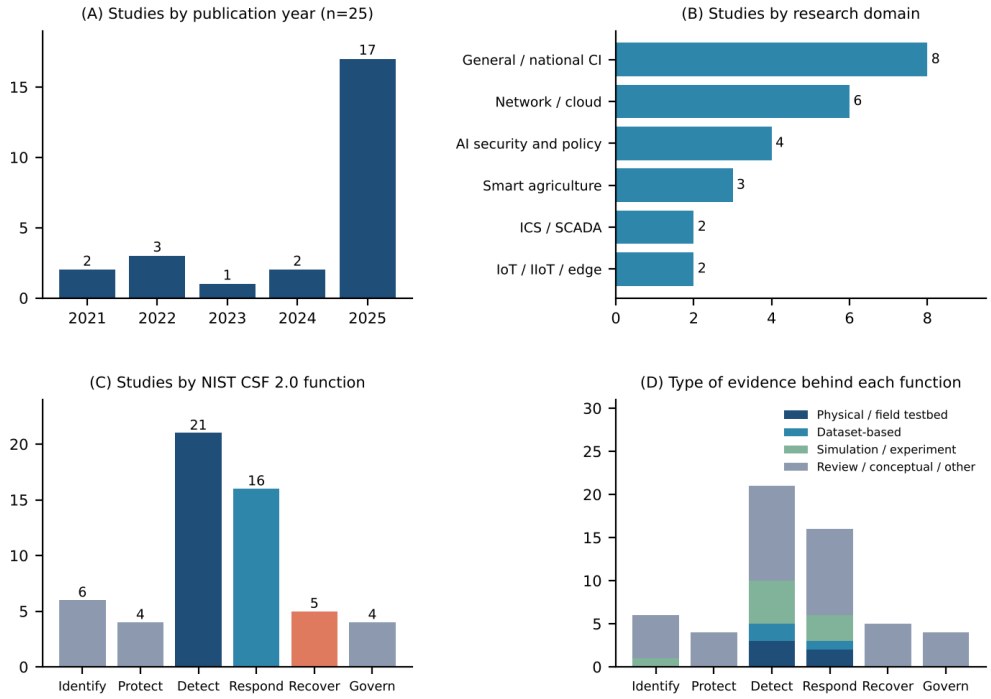


Fig. 5. (A) Distribution of the 25 studies by publication year. (B) Distribution by research domain. (C) Number of studies addressing each NIST CSF 2.0 function; Recover has minimal coverage. (D) Type of evidence behind each function, as classified in Table 5 (a study is counted once for every function it addresses).

7.8 Developing-Region and Pakistan Context

The only field-grounded agentic agriculture study in the corpus comes from Pakistan: Murad and Ahmed [16] collected 12,000 sensor samples in Karachi and Thatta and deployed a multi-agent system with a field robot, yet no study in the corpus evaluates the cyber security of such a system. This matters because the conditions that make agentic agriculture attractive in developing regions (intermittent connectivity, low-cost sensors, regional-language interfaces and limited security staff) also make it hard to defend. Three directions follow from the evidence reviewed here, not from region-specific incident data, which the corpus does not contain: edge-resident detection that keeps working when the cloud is unreachable, building on [16] and [24]; lightweight federated detectors suited to constrained hardware, building on [17]; and national or sectoral NIST CSF 2.0 profiles for farming, as argued in [9] and [21]. Regional incident data would be needed to turn these directions into tested recommendations.

8 Conclusion and Future Directions

Agentic AI for critical-infrastructure cybersecurity has moved beyond proof of concept. Across the 25 reviewed studies, systems built on reinforcement learning, LLM reasoning, edge-based federated learning and multi-agent orchestration reach accuracy and latency well beyond traditional defences. Where evaluation conditions are comparable, detection accuracy on heterogeneous data sits at 93 to 99 percent, RL-based designs respond in under a second, and federated learning at the edge adds only about

10 percent computational overhead for privacy protection [17]. These figures come from the source studies, are subject to the comparability limits discussed in Sections 5 to 7, and mark the current upper envelope of the field rather than guaranteed deployment performance. Markets are moving the same way: agentic cybersecurity is projected to reach 146.5 billion dollars by 2034 [2], and 67 percent of Fortune 500 companies planned to adopt agentic AI by the end of 2025 [30].

The evidence is uneven. Detection is well studied, while recovery and self-healing are in their infancy. ICS have a solid empirical base, but agricultural IoT lags despite comparable strategic risk. Governance work has begun but trails deployment [19, 30]. Most worrying, the autonomy that makes these systems useful also makes them targets. In [6] detection collapses to zero when the defensive layer is removed, and the agentic threat taxonomy of [19] shows how reasoning, memory, tool use and multi-agent interaction each add attack surface. Measured rates for such attacks against deployed defenders are largely missing from the literature, which is itself a gap.

We recommend three priorities. First, build a shared benchmark for agentic ICS defence that uses real industrial protocols and physical PLCs, extending the Houkan dataset [7] into a community resource with standardised scenarios. Second, invest in agricultural IoT security at a level that matches the importance of food production, with field trials like those of Murad and Ahmed [16]. Third, treat agentic AI governance as a research track on equal footing with detection and response, grounded in NIST CSF 2.0 [19], ISO/IEC standards and the Principle of Least Agency. The technology will be deployed. The open question is whether the institutions around it are ready to deploy it safely.

Answers to the research questions.

RQ1: the literature falls into LLM-based, RL-based, federated or edge, multi-agent and hybrid families, concentrated on Detect (21 of 25 studies) and Respond (16), with Identify, Protect, Govern and Recover each covered by 6 or fewer (Sections 4 and 5, Fig. 2).

RQ2: reported detection accuracy clusters at 93 to 99 percent and RL decisions take milliseconds, whereas LLM agents are about 104 times slower; 8 of the 11 experimentally validated studies rely on simulation or recorded datasets, so figures are upper bounds (Sections 6.3 and 7, Table 5).

RQ3: agentic defenders add attack surface through reasoning, memory, tool use and multi-agent interaction [19], few studies measure these risks, and governance proposals converge on least agency, escalation thresholds and auditability [19, 30] (Sections 6.4 and 7.4).

RQ4: the priorities are a shared physical-testbed benchmark for agentic ICS defence, field-validated agricultural IoT security, and recovery with governance (Sections 5.1 and 7.8).

References

[1] Yigit, Y. and M.A. Ferrag, *Generative AI and LLMs for critical infrastructure protection: evaluation benchmarks, agentic AI, challenges, and opportunities.* Sensors, 2025. **25**(6): p. 1666.

[2] Kshetri, N., *Transforming cybersecurity with agentic AI to combat emerging cyber threats.* Telecommunications Policy, 2025. **49**(5): p. 102976.

[3] Karamchand, G. and O.O. Aramide, *AI and cyberwarfare.* J. Tianjin Univ. Science and Technology, 2025. **58**(8).

[4] Hanif, M.A., *Bringing autonomy and cooperation together: a comparison of agentic AI systems and AI agents.* ACM Computing Surveys, 2025. **57**(4): p. 1–38.

[5] Kott, A., *Autonomous Intelligent Cyber defense Agent: introduction and overview*, in *Autonomous Intelligent Cyber Defense Agent (AICA)*. 2023, Springer: Cham. p. 1–25.

[6] Shrestha, S., *Investigation of cybersecurity bottlenecks of AI agents in industrial automation.* Computers, 2025. **14**(11): p. 456.

[7] Houkan, A., et al., *Artificial intelligence approach to intrusion detection in industrial control systems with real world dataset generation and model evaluation.* Discover Artificial Intelligence, 2025. **5**: p. 307.

[8] Rajgopal, P.R., *Agentic AI for autonomous cybersecurity threat mitigation.* Journal of Cybersecurity and Privacy, 2025. **5**(2): p. 234–251.

[9] Campoverde Molina, M. and S. Lujan Mora, *Cybersecurity in smart agriculture: a systematic literature review*. Computers & Security, 2025. **150**: p. 104284.

[10] Sebastian, R. *Large language models are autonomous cyber defenders*. in *Proceedings of the IEEE Conference on Communications and Network Security*. 2025.

[11] Wolk, M., *Beyond CAGE: investigating generalization of learned autonomous network defense policies.* arXiv preprint arXiv:2211.15557, 2022.

[12] Collyer, J., A. Andrew, and D. Hodges. *ACD G: enhancing autonomous cyber defense agent generalization through graph embedded network representation*. in *Proc. of the 39th ICML (ML4Cyber Workshop)*. 2022.

[13] Prebot, B., et al. *Cognitive models of dynamic decision in autonomous intelligent cyber defense*. in *Proceedings of the 2nd International Workshop on Adaptive Cyber Defense*. 2022.

[14] Tran, K. and A. Akella, *Deep hierarchical reinforcement agents for automated penetration test.* arXiv preprint arXiv:2109.06449, 2021.

[15] Mitchell, J.A., R.L. Cater, and M.D. Hernandez, *Agentic AI frameworks for autonomous threat detection in national critical infrastructure.* International Journal of Critical Infrastructure Protection, 2025. **49**: p. 100712.

[16] Murad, M. and M. Ahmed, *Agentic AI framework to automate traditional farming for smart agriculture.* AgriEngineering, 2025. **8**.

[17] Toree, D., *Toward enhancing privacy preservation of a federated learning CNN intrusion detection system in IoT.* ACM Trans. Software Engineering & Methodology, 2025. **34**(2): p. Art. 53.

[18] Adabara, I., *A review of agentic AI in cybersecurity: cognitive autonomy, ethical governance, and quantum resilient defense.* F1000Research, 2025. **14**: p. 843.

[19] Metin, B., et al., *Securing agentic AI with NIST cybersecurity framework.* ELECO 2025 Int. Conference on Electrical & Electronics Eng.. 2025.

[20] Ibitoye, J.S., *Self healing AI driven networks for automated cyber threat detection and recovery.* Global J. Engineering and Tech. Advances (GJETA), 2021. **9**(3): p. 154–169.

[21] Alahe, M.A., *Cybersecurity in smart agriculture: threat types, current status, and future trends.* Computers and Electronics in Agriculture, 2024. **226**: p. 109401.

[22] Gvilava, N.T., *Autonomous intelligent agents in decision support systems for critical infrastructure.* Critical Infrastructure Protection Review, 2025. **11**(3): p. 215–232.

[23] Bondhala, S., *Cybersecurity in AI driven data centers: reinventing threat detection.* IJARSCT, 2025. **5**(7): p. 510–519.

[24] Ajayi, R. and M. Masunda, *Integrating edge computing, data science and advanced cyber defense for autonomous threat mitigation.* Int. J. Science and Research Archive, 2025. **15**(2): p. 063–080.

[25] Hussain, A., H. Sharif, F. Rehman, H. Kirn, A. Sadiq, M.S. Khan, A. Riaz, C.N. Ali, et al., *A systematic review of intrusion detection systems in Internet of Things using ML and DL.* 2023 4th International Conference on Computing, Mathematics and Engineering Technologies, 2023.

[26] Malik, K., F. Rehman, T. Maqsood, S. Mustafa, O. Khalid, and A. Akhunzada, *Lightweight internet of things botnet detection using one-class classification.* Sensors, 2022. **22**(10): p. 3646.

[27] Rehman, F., A. Rabail, M.H. Sajjad, and S. Gul, *AI integration in drone technology: revolutionizing applications in agriculture, security, and beyond.* In Unmanned Aerial Vehicles Swarm for Protecting Smart Cities: Future Trends and Applications, 2025.

[28] Muneer, M., F. Rehman, M.H. Sajjad, M. Anwar, and K.N. Qureshi, *Security and privacy concerns in AI models.* In Next Generation AI Language Models in Research, 2024: p. 293-326.

[29] Shaukat, S., F. Rehman, G. Chhabra, K. Kaushik, I.U. Khan, H. Sharef, M.Z. Ali, et al., *Examination of phishing attempts on web-based business applications and their preventions.* 2024 International Conference on Communication, Computer Sciences and Engineering, 2024.

[30] Joshi, S., *AI governance in the era of agentic generative AI and AGI: frameworks, risks, and policy directions.* IJIRCST, 2025. **13**(5): p. 14–24.

[31] Rizwan Hameed A COMPREHENSIVE REVIEW OF INTRUSION DETECTION SYSTEMS ACROSS DATASET FAMILIES", Kashf J. Multidiscip. Res., vol. 3, no. 01, pp. 79–92, Jan. 2026, doi: 10.71146/kjmr814.

[32] Hameed, R., *ADNN-FER: adaptive deep neural networks for real-time facial expression recognition using hybrid attention, compound loss functions, and transformer-enhanced feature fusion.* J. Artif. Intell. Mach. Learn. Neural Netw., 2025. **5**(2): p. 128–137.