Review | Open Access

Intelligence & Robotics

Huang *et al. Intell. Robot.* **2026**, *6*(4), 919-69 **DOI:**10.20517/ir.2026.40

Secure control of multi-agent systems under cyber attacks: a comprehensive survey

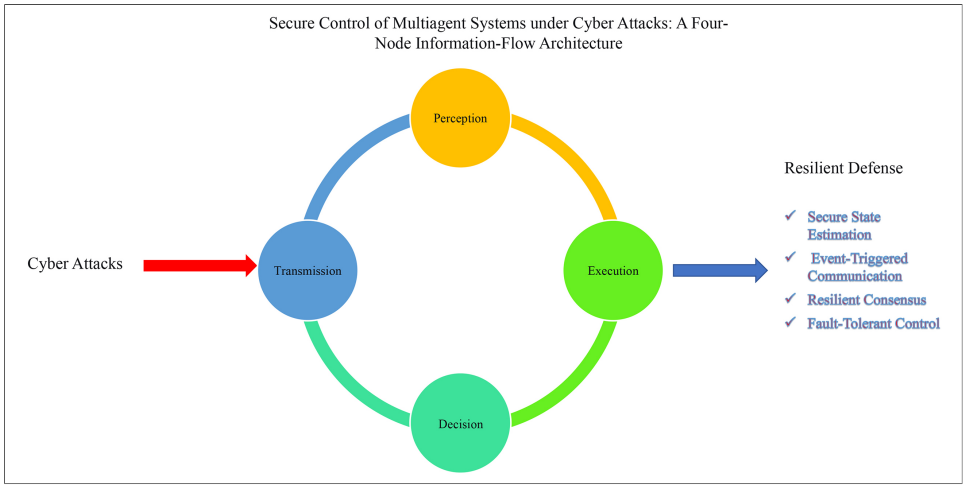
Kaijiao Huang[1,2], **Baiming Shi**[1], **Jian Chen**[1]

Keywords:
Multi-agent systems, cyber-physical security, secure control, event-triggered control, information-flow architecture

Citation: Huang, K.; Shi, B.; Chen, J. Secure control of multi-agent systems under cyber attacks: a comprehensive survey. *Intell. Robot.* **2026**, *6*(4), 919-69. https://dx.doi.org/10.20517/ir.2026.40

Received: 1 Aug 2026 **First Decision:** 21 Sep 2026 **Revised:** 23 Oct 2026 **Accepted:** 27 Oct 2026 **Published:** 11 Nov 2026

Academic Editor: Chen Peng **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang



Abstract

Multi-agent systems (MASs) rely on open communication networks and are thus exposed to malicious cyber attacks, which not only threaten system performance but also erode the fundamental basis of collaborative operation. This paper takes the novel four-node information flow architecture as the main line, divides the cooperative control loop into four functional layers: perception, transmission, decision-making, and execution, and conducts a systematic and comprehensive review of the latest research progress in the security control of MASs. Under this architecture, we establish a unified attack classification system that captures the mathematical commonalities and operational characteristics of denial-of-service, spoofing, Byzantine, and hybrid attacks. Subsequently, we systematically sorted out the defense strategies along four information flow nodes: the perception layer covers security state estimation and physical layer trust; the transmission layer involves event-triggered, self-triggered, and quantized communication; the decision-making layer includes resilient consensus, observer-based control, adaptive control, and reinforcement learning control; the execution layer focuses on fault-tolerant control, sliding mode control, and impulsive control. A core theme running through the entire text is the safety-performance-resource tripartite paradox that restricts the design of elastic MASs. We examined the path of evolution from passive robust defense to active defense to

Check for updates

[1]College of Engineering, China Agricultural University, Beijing 100083, China.
[2]School of Mathematical Sciences & Center for Applied Mathematics of Guangxi, Guangxi Minzu University, Nanning 530006, Guangxi, China.

Correspondence to: Prof. Jian Chen, College of Engineering, China Agricultural University, Beijing 100083, China. E-mail: jchen@cau.edu.cn

reconcile these conflicting goals and pointed out open challenges such as cross-layer collaborative design, data-driven resilience, asynchronous multi-channel attacks, and verification on actual platforms. This survey also highlights emerging application fields, such as clusters of intelligent agricultural robots. Under adversarial communication conditions, secure collaboration is crucial for precision agriculture, autonomous harvesting, and field monitoring. By taking a unified perspective that integrates attack modeling and defense mechanisms, this survey aims to provide researchers and engineering practitioners with a systematic and forward-looking reference blueprint for building intrinsically secure collaborative MASs.

1. INTRODUCTION

Multi-agent systems (MASs) consist of spatially distributed agents, such as unmanned aerial vehicles, mobile robots, autonomous underwater vehicles, and smart grid substations, that collaborate through communication networks[1-5]. They accomplish complex cooperative tasks through local information exchange, making them a highly influential technical paradigm. Although a single agent possesses limited sensing, computing, and actuation capabilities, the swarm intelligence emerging from networked interactions enables the system to exceed the sum of individual capabilities.

This structural advantage has driven the widespread deployment of MASs across critical infrastructure domains, including unmanned aerial vehicle (UAV) swarm formation[2,6], satellite attitude coordination[3], multi-robot collaboration[4,7], distributed detection over sensor networks[5,8], microgrid energy management[9], and agricultural embodied intelligent robot swarm networks[10]. Agricultural embodied intelligent robot swarms are not merely an application background; they drive new security and control problems that generic multi-agent formulations do not capture. Ground machinery for seeding, spraying, and harvesting exhibits Euler–Lagrange (EL) dynamics with nonholonomic constraints, while UAVs for crop monitoring and mapping are underactuated, strongly nonlinear, and aerodynamically coupled. These platforms differ in state dimension, actuation authority, communication range, and vulnerability profile. Their collaboration is both informational and physical: UAVs provide overhead perception and communication relaying, while ground robots execute physical tasks. This physical–informational coupling creates attack surfaces absent in homogeneous MASs. A compromised UAV can inject false crop-health maps into the ground network, causing incorrect spraying decisions; a compromised ground robot can broadcast false task-completion messages, misleading UAV scheduling. Moreover, these robots rely on multimodal field sensing, global navigation satellite system (GNSS)/real-time kinematic (RTK), stereo cameras, light detection and ranging (LiDAR), inertial measurement unit (IMU), and multispectral imagers, each with distinct deception modes: GNSS spoofing falsifies positions, camera spoofing imitates crop rows, multispectral deception injects false normalized difference vegetation index (NDVI) values, and LiDAR/IMU replay misleads navigation. Farmland communication channels are equally hostile, with intermittent connectivity, long-range loss, terrain and canopy obstruction, dust, rain, and high temperature causing severe fading, delay, and packet loss. Secure control must therefore address not only consensus and formation stability but also task-level semantic consistency and physical safety constraints such as sprayer pressure limits, weeder torque bounds, and obstacle avoidance margins.

The fundamental difference between MASs and traditional centralized control systems lies in their information-driven decision-making paradigm: the control actions of each agent rely on the real-time reception and processing of state information from neighboring agents rather than on the aggregation and central processing of global information. This structural feature implies that cooperative control correctness fundamentally depends on the integrity and timeliness of the information flow carried by the communication network. However, the openness of wireless channels, the broadcast nature of data transmission, and the widespread heterogeneity in large-scale deployments together create a large attack

surface. Once the communication network is damaged, the information foundation supporting collaborative decision-making collapses, and the system becomes highly prone to catastrophic failure. Major security incidents have repeatedly confirmed this vulnerability. The Stuxnet worm paralyzed Iranian centrifuges; attackers remotely invaded a water plant's supervisory control and data acquisition (SCADA) system, causing a pump to burn out; and the Havex virus shut down hydropower stations and overloaded nuclear power plants[11]. As MASs increasingly manage security-critical infrastructure, successful attacks can have consequences far beyond economic losses; they directly threaten national security, environmental security, and human life.

From a control theory perspective, cyber attacks pose threats fundamentally different from traditional uncertainties such as process noise or model errors. First, the attack surface expands exponentially: as the system grows from a single agent to N agents, the number of potential attack entry points multiplies[12,13]. Second, information coupling causes cascading fault propagation: within a few rounds of information exchange, compromising a single agent can contaminate the entire network[14-16]. Third, MASs incorporate multiple control objectives such as consensus, formation, containment, and optimization, and the sensitivity of different objectives to information integrity varies significantly. This means that the same attack may affect different control tasks in markedly different ways[17-19]. Therefore, cyberattacks threaten not merely system performance but the very foundation of collaboration itself[20,21]. This understanding has triggered a paradigm shift from traditional collaborative control to secure collaborative control. Traditional cooperative control implicitly assumes trustworthy communication; in contrast, secure collaboration regards network uncertainties such as communication interruption, data tampering, and node betrayal as inherent system attributes that must be considered from the very beginning of controller design[22]. Security is thus becoming an endogenous dimension of multi-agent control theory, just as the immune system of an organism is not an external add-on but an organic component deeply integrated into its life activities.

Among the various cyber threats confronting MASs, denial-of-service (DoS) attacks and deception attacks represent the two most prevalent and destructive categories. DoS attacks block information transmission by overwhelming communication channels or jamming signals, preventing agents from exchanging critical state information and causing intermittent connectivity, time-varying topologies, and even complete isolation of nodes[23,24]. Deception attacks, such as false data injection (FDI), compromise data integrity by tampering with transmitted signals, leading agents to make incorrect decisions based on falsified information[25]. The insidious nature of deception attacks lies in their stealth: the system may appear to function normally while gradually diverging from its intended behavior. In practice, these two attack types may occur simultaneously, forming hybrid attack modes that pose even greater challenges for secure control design[26]. Early studies often follow the seminal framework of De Persis and Tesi[24], imposing constraints on attack frequency and duration. More recent works have introduced stochastic models, such as two-mode Markov processes, to capture random activation patterns. Moreover, adversaries may launch asynchronous multi-channel attacks that independently target specific communication links[27,28], or even compromise individual agents to act as Byzantine nodes that deliberately disseminate misleading information[20]. In agricultural embodied intelligent robot swarms, DoS attacks are particularly damaging because rural areas lack redundant communication infrastructure, while FDI attacks may target GPS receivers or multispectral sensors used for crop health assessment. An attacker could spoof the position of a harvesting robot or inject false NDVI readings, leading to incorrect actuation decisions and potentially devastating crop losses.

To address these threats, researchers have developed a variety of resilient control strategies. A prominent approach is to design distributed observers that reconstruct the state of leaders and system dynamics even under intermittent communication failures, handling delay and non-uniform sampling through adaptive and sampled-data mechanisms[29,30]. Event-triggered and self-triggered communication schemes reduce

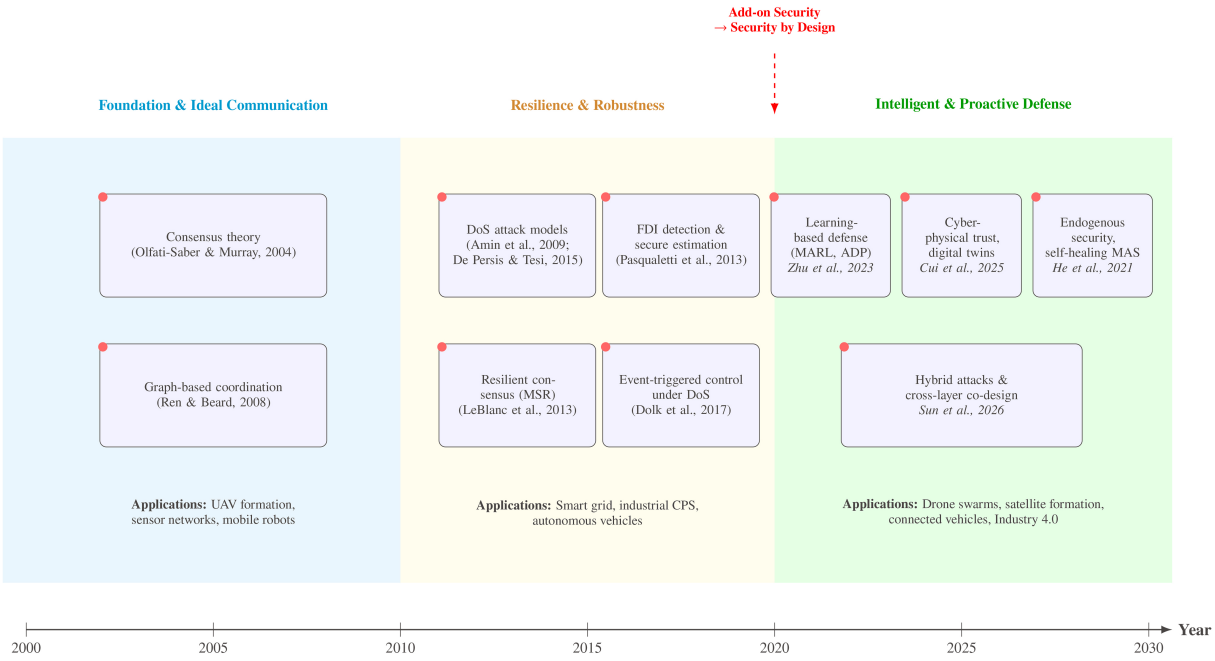


Figure 1. Technological evolution of secure control for MASs. The representative milestones are compiled from the works[14,24,27,50-57]. MASs: Multi-agent systems; UAV: unmanned aerial vehicle; DoS: denial-of-service; FDI: false data injection; MSR: mean-subsequence-reduced; CPS: cyber-physical systems; MARL: multi-agent reinforcement learning; ADP: adaptive dynamic programming.

communication burdens while maintaining system performance under attacks: an agent transmits information only when a predefined triggering condition is violated, thereby reducing exposure to opportunistic attacks and saving bandwidth and energy[27,31-33]. Impulsive control applies discrete corrections at specific moments, combines robustness with rapid response, and can be combined with topology switching to achieve secure synchronization under hybrid attacks[34]. Adaptive fuzzy control with neural networks (NNs) simultaneously compensates for unknown nonlinearities and actuator faults, and can handle DoS attacks as well[31,35-37]. To balance stability and performance under resource constraints, reinforcement learning (RL) and model-free adaptive predictive control have emerged; they complete control synthesis relying solely on input and output data without an explicit system model[38]. Moreover, finite-time, fixed-time, and predefined-time control schemes guarantee convergence within a user-specified time window even under malicious interference[39-41]. In addition to consensus and formation control, distributed optimization and economic scheduling under DoS attacks have attracted widespread attention, with algorithms capable of handling coupled cost functions and equality constraints[42]. More recently, innovative defense paradigms have been proposed, such as digital twin-assisted frameworks[43], physical-layer trust measurement based on radio frequency (RF) fingerprints and channel state information[44], and multi-agent reinforcement learning (MARL) for autonomous detection and mitigation of network threats[45].

Several comprehensive reviews have been published on related topics, covering attack detection in cyber-physical systems[46], security control and attack detection in industrial cyber-physical systems[47], defense strategies against disclosure/deception/sabotage attacks[48], DoS and FDI attack models in smart grids[17], deterministic and probabilistic DoS models in networked control systems[49], and secure state estimation (SSE) and control in cyber-physical systems[21]. Figure 1 sketches the technological evolution of security control in MASs, from the theoretical foundation period under ideal communication conditions to the emerging stage of intelligent active defense. Table 1 compares existing surveys with this article in detail.

Table 1. Comparison of representative surveys on security of cyber-physical systems and MASs

Ref.	Primary focus	Attack types	System dynamics	Control and defense methods	Unified attack framework	Defense taxonomy based on information flow	Resource-aware
[46]	Attack detection in CPS	DoS, deception	Linear, nonlinear	Detection, estimation	×	×	×
[47]	Security control and detection for industrial CPS	DoS, deception	Linear, nonlinear	Control, detection	×	×	×
[48]	Defense strategies for disclosure, deception, sabotage	Disclosure, deception, sabotage	Linear	Robust control, detection	×	×	×
[17]	DoS and FDI in smart grids	DoS, FDI	Power system models	State estimation, detection	×	×	×
[49]	DoS attack models in NCS	DoS	Linear, nonlinear	Control under packet loss	×	×	×
[21]	SSE and control of CPS	DoS, deception	Linear, nonlinear	Observer-based control, estimation	×	×	×
[20]	Secure control of MASs against malicious attacks	DoS, deception, Byzantine	Linear, nonlinear	Resilient consensus, detection	×	×	×
[44]	Physical-layer trust for CPS	Spoofing, jamming	Generic	Trust mechanisms, signal fingerprinting	×	×	×
[45]	MARL for cybersecurity	Generic (policy attacks)	Generic	RL	×	×	×
This survey	Secure control of MASs under cyber attacks	DoS, deception, Byzantine, asynchronous multi-channel, hybrid	Linear, nonlinear, heterogeneous, fractional-order, Euler-Lagrange	Event-triggered, observer-based, adaptive, RL, impulsive, fixed-time, model-free, switching	√	√	√

√: Explicitly addressed with a novel framework; ×: not addressed or only marginally mentioned. MASs: Multi-agent systems; CPS: cyber-physical systems; DoS: denial-of-service; FDI: false data injection; NCS: networked control systems; SSE: secure state estimation; MARL: multi-agent reinforcement learning; RL: reinforcement learning.

Two closely related lines of work deserve explicit comparison because they address attack classes and system settings central to this survey. The first is secure event-triggered consensus control of linear MASs subject to sequential scaling attacks proposed by He and Mo[58]. In that work, a scaling deception attack is modeled as a sequential operator that scales the transmitted state or control signal during specific time intervals, characterized by explicit attack duration and frequency constraints rather than probabilistic models. The authors design static and dynamic event-triggered control protocols, derive sufficient consensus conditions relating attack duration and frequency to triggering parameters, and prove Zeno-freeness. Within the four-node information-flow architecture, this work acts primarily at the transmission node by reducing communication exposure through event-triggering, and partially at the decision node by adapting the consensus protocol to scaling attacks. Its attack model is a special case of multiplicative FDI on communication links, and its defense strategy is a cross-layer transmission–decision mechanism. However, it does not address Byzantine agents, replay attacks, actuator attacks, or the perception layer, and its stability guarantee assumes linear agent dynamics and a known attack structure. The second closely related work is the resilient vector consensus over random dynamic networks under mobile malicious attacks. The literature[59] investigates resilient vector consensus for dynamic agents against mobile malicious attacks, where the adversary can move across the network over time and compromise different agents. Three mobile attack models are considered, and distributed discrete-time algorithms are developed using Lyapunov theory

and martingale convergence. Resilient vector consensus is shown to be reached under stochastic robustness conditions on the underlying random dynamic network. In the four-node framework, this work focuses on the decision node and develops resilient consensus algorithms that tolerate Byzantine-like or malicious nodes. Its network model is random and dynamic, more general than static or deterministically switching topologies, and its mathematical tools - martingale convergence and stochastic robustness conditions - differ fundamentally from the switched-system and dwell-time analysis used for DoS resilience. However, it does not address communication-layer DoS, FDI on links, or actuator-level attacks, and its guarantee is asymptotic in expectation rather than deterministic. These two works are complementary in attack class, network model, control objective, and mathematical machinery. Neither integrates perception-layer secure estimation, actuator-layer fault tolerance, or resource-aware cross-layer co-design. The present survey incorporates both lines of work into the four-node information-flow framework, generalizes their attack models into a unified taxonomy, and extends the defense taxonomy to cover perception, transmission, decision-making, and execution as coupled nodes.

This survey proposes a unified classification system that organizes existing achievements along three dimensions: attack modeling, defense mechanisms, and structural tension. Attack modeling formally describes DoS, FDI, Byzantine, and hybrid attacks, capturing both their mathematical commonalities and operational characteristics. Defense mechanisms are systematically classified according to a four-node information-flow architecture comprising perception, transmission, decision-making, and execution, which reveals the functional complementarity among different defense strategies and explicitly distinguishes single-layer from cross-layer mechanisms. Structural tension refers to the critical examination of the security–performance–resource trilemma, the gap between model fidelity and actual deployability, and the paradigm shift from passive robustness to active defense. As a supplement to this framework, we critically review key methodological innovations such as observer compensation, event-triggered protocols, Lyapunov-based stability analysis, and switching-system characterization. Throughout the review, we consistently highlight the fundamental practical tensions that shape this field, including the trilateral paradox of security, performance, and resources, and the gap between model fidelity and actual deployability, as well as the paradigm shift from passive resilience to active defense.

The novelty of the proposed four-node information-flow architecture lies in its feedback-loop-centric organization, which differs fundamentally from existing cyber-physical systems (CPS)/MAS security taxonomies. Existing classifications are typically organized by attack type, such as DoS, false injection, replay, and Byzantine attacks[47,48], by defense tool, such as observer-based detection, event-triggered control, and RL[21,46], or by architecture layer, such as the device, network, control, data, algorithm, and application layers in the Architecture-Driven Hexa-Layer (ADHL) taxonomy[60] and the seven-layer attack surface model for agentic artificial intelligence (AI) systems[61]. Although useful for cataloging the literature, these taxonomies treat attacks and defenses as largely independent categories and do not explicitly reveal how a defense acts within the closed-loop information flow, nor how different layers interact. The four-node architecture instead decomposes the cooperative control loop into perception, transmission, decision-making, and execution, and treats these nodes as coupled components of a single feedback system. This provides three distinctive advantages: it maps each attack to the node where it corrupts information and each defense to the node where it restores availability, integrity, or physical safety; it explicitly distinguishes single-layer defenses from cross-layer mechanisms such as event-triggered observer-based resilient control[62]; and it exposes the security–performance–resource trilemma at each node, enabling systematic comparison of assumptions, stability guarantees, communication burden, and computation cost. Existing taxonomies answer “what is the attacker doing?” or “which tool is used?”; the four-node architecture answers “where in the feedback loop does the defense act, and how does it couple with other nodes?” This complementary perspective is what

enables the cross-layer coupling analysis and resource-aware defense taxonomy developed throughout this survey.

To ensure transparency and reproducibility, this survey adopts a systematic literature search and selection process. We conducted the initial systematic search in IEEE Xplore, Web of Science, Scopus, ScienceDirect, SpringerLink, and arXiv for publications from January 2010 to December 2025, including seminal earlier works when necessary. During the revision stage, this literature base were supplemented with several cutting-edge publications that were accepted or published online in early 2026 (e.g.,[27]) to ensure the survey captures the most recent advances. The search combined three groups of keywords: system and application terms (e.g., MAS, cyber-physical system, cooperative control, robot swarm, UAV formation, agricultural robot, and embodied intelligence); security and attack terms (e.g., secure control, cyberattack, DoS, FDI, replay attack, Byzantine attack, spoofing, actuator attack, and hybrid attack); and defense and methodology terms (e.g., resilient consensus, event-triggered control, self-triggered control, quantized communication, observer-based control, adaptive control, RL, federated learning, fault-tolerant control, sliding mode control (SMC), impulsive control, and privacy-preserving control), combined with Boolean operators. Publications were included if they were peer-reviewed journal articles, peer-reviewed conference papers, or widely cited preprints; addressed security, resilience, or privacy in multi-agent, cyber-physical, or networked control systems; provided theoretical analysis, algorithm design, experimental validation, or a systematic survey; and were directly relevant to at least one of the four information-flow nodes, namely perception, transmission, decision-making, or execution. The studies were excluded if they focused exclusively on cryptographic primitives, authentication protocols, or network-layer security without control-loop or physical-dynamics relevance; addressed general cybersecurity or machine learning without a clear multi-agent or cyber-physical control setting; duplicated existing contributions without new technical content; or were not written in English. The initial search returned approximately 1,200 records; after duplicate removal, approximately 800 titles and abstracts were screened, approximately 350 full texts were assessed against the inclusion and exclusion criteria, and approximately 188 references were retained. For each included publication, we extracted the attack model, system dynamics, control objective, defense mechanism, stability or security guarantee, communication and computation requirements, validation method, and application domain. These attributes underpin the comparative tables and the four-node information-flow taxonomy used throughout this survey, ensuring balanced and representative coverage of the literature.

The remainder of this survey is structured as follows. Section 2 establishes the mathematical foundations, including algebraic graph theory, MAS dynamics, a unified attack taxonomy, and the four-node information-flow architecture that serves as the conceptual backbone of this work. Section 3 systematically reviews defense mechanisms across the four information-flow nodes: perception-layer defense, transmission-layer defense, learning and adaptive control, and execution-layer defense. Section 4 outlines ten interconnected future research directions that form a coherent progression from distributed intelligent defense to proactive AI-powered security. Section 5 concludes the survey by synthesizing key insights and articulating the paradigm shift from add-on security to endogenous resilience.

2. MATHEMATICAL FOUNDATIONS AND A UNIFIED ATTACK–DEFENSE MODELING FRAMEWORK FOR SECURE MASS

2.1. Graph theory and communication topologies

In MASs, the information exchange among agents can naturally be modeled with the aid of algebraic graph theory. Depending on the communication direction, the interaction topology can be represented as either an undirected or a directed graph. Furthermore, as network attacks such as DoS or spoofing attacks may block communication links, network topologies often exhibit time-varying or switching characteristics[63]. Therefore, it is necessary to extend classical graph theory concepts to dynamic scenarios.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ denote a weighted graph[64], where $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$ is the set of nodes (agents), $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges (communication links), and $\mathcal{A} = [a_{ij}] \in \mathbb{R}^{N \times N}$ is the adjacency matrix. An edge $(v_j, v_i) \in \mathcal{E}$ indicates that agent i can receive information from agent j, with weight $a_{ij} > 0$; otherwise, $a_{ij} = 0$. We assume no self-loops, i.e., $a_{ii} = 0$ for all i. The neighbor set of agent i is defined as $\mathcal{N}_i = \{j \in \mathcal{V} \mid (v_j, v_i) \in \mathcal{E}\}$.

For a directed graph (digraph), the in-degree of node i is $d_{\text{in},i} = \Sigma_{j=1}{}^{N} a_{ij}$ and the out-degree is $d_{\text{out},i} = \Sigma_{j=1}{}^{N} a_{ji}$. The Laplacian matrix $\mathcal{L} = [l_{ij}] \in \mathbb{R}^{N \times N}$ is defined as

$$l_{ij} = \begin{cases} -a_{ij}, & i \neq j, \\ \sum_{k=1, k \neq i}^{N} a_{ik}, & i = j. \end{cases} \tag{1}$$

Equivalently, $\mathcal{L} = \mathcal{D}_{\text{in}} - \mathcal{A}$, where $\mathcal{D}_{\text{in}} = \text{diag}\{d_{\text{in},1}, \ldots, d_{\text{in},N}\}$.

For an undirected graph, $(v_j, v_i) \in \mathcal{E}$ implies $(v_i\, v_j) \in \mathcal{E}$ and $a_{ij} = a_{ji}$. In this case, $\mathcal{A}$ is symmetric, the degree of node i is $d_i = \Sigma_{j=1}{}^{N} a_{ij}$, and the Laplacian matrix is symmetric and positive semidefinite:

$$\mathcal{L} = \mathcal{D} - \mathcal{A}, \tag{2}$$

where $\mathcal{D} = \text{diag}\{d_1, \ldots, d_N\}$. The Laplacian $\mathcal{L}$ has a zero eigenvalue with associated eigenvector $\mathbf{1}_N = [1, \ldots, 1]^{\top}$, and all other eigenvalues are positive if and only if the graph is connected[64,65].

A directed graph contains a spanning tree if there exists a root node from which there is a directed path to every other node. For a digraph, $\mathcal{L}$ has at least one zero eigenvalue, and all non-zero eigenvalues have positive real parts if and only if the graph contains a spanning tree[50,51].

In addition to deterministic or switching topologies, recent work has considered random dynamic networks under mobile malicious attacks[59]. In this setting, the network topology evolves according to a stochastic process, and the set of malicious nodes changes over time as the adversary moves across the network. The graph-theoretic robustness conditions are then expressed in terms of stochastic robustness rather than deterministic connectivity or spanning-tree conditions. This motivates the inclusion of stochastic robustness conditions alongside the union-graph and spanning-tree conditions discussed above, particularly for decision-layer resilient consensus against mobile Byzantine agents.

In many cooperative control tasks such as formation control and output regulation, a subset of agents, called leaders, generate reference trajectories that the remaining agents, called followers, must track. Consider an MAS with M followers and N - M leaders. The communication graph among the followers is denoted by $\mathcal{G}_F$, with Laplacian matrix $\mathcal{L}_F$. The interaction from leaders to followers is captured by a diagonal matrix $\mathcal{B} = \text{diag}\{b_1, \ldots, b_M\}$, where $b_i > 0$ if follower i receives information from at least one leader, and $b_i = 0$ otherwise. The information exchange among all agents can then be described by the overall Laplacian-like matrix $\mathcal{H} = \mathcal{L}_F + \mathcal{B}$, which plays a central role in the stability analysis of leader–follower consensus[51,66].

Under DoS attacks or other communication impairments, the set of active links becomes time-dependent. We model the attacked communication network as a time-varying graph $\mathcal{G}(t) = (\mathcal{V}, \mathcal{E}(t), \mathcal{A}(t))$, where the edge set $\mathcal{E}(t)$ and adjacency weights $a_{ij}(t)$ are piecewise constant and switch at isolated time instants[24,67]. Specifically, let $\{\tau_k\}_{k \in \mathbb{N}}$ be the switching instants and let $\mathcal{G}_{\sigma(t)}$ denote the active topology, where σ: $[0, \infty) \to \mathcal{P}$ is a piecewise constant switching signal taking values in a finite index set $\mathcal{P}$. The Laplacian matrix at time t is then $\mathcal{L}_{\sigma(t)}$.

To ensure consensus or other cooperative objectives despite attacks, one often requires that the union graph satisfies connectivity conditions over certain time intervals. For a collection of graphs $\{\mathcal{G}_1, \ldots, \mathcal{G}_m\}$ on the same vertex set, the union graph is defined as $\bigcup_{k=1}^{m} \mathcal{G}_k = (\mathcal{V}, \bigcup_{k=1}^{m} \mathcal{E}_k)$. A common sufficient condition for consensus under switching topologies is the existence of an infinite sequence of bounded intervals such that the union graph over each interval contains a spanning tree[67].

Definition 1[14]**.** In the context of Byzantine attacks, the adversary can compromise a subset of agents. Let $\mathcal{F} \subset \mathcal{V}$ be the set of malicious nodes. The system is said to be under an F-local attack if for every normal agent i, at most F of its neighbors are malicious, i.e., $|\mathcal{N}_i \cap \mathcal{F}| \leq F$. An F-total attack is one in which the total number of malicious agents in the network is at most F, i.e., $|\mathcal{F}| \leq F$.

2.2. MAS dynamics

We consider an MAS consisting of N autonomous agents. The dynamics of agent i ($i = 1, 2, \ldots, N$) can be described by a general nonlinear ordinary differential equation coupled with an output equation:

$$\begin{aligned} \dot{x}_i(t) &= f_i\big(x_i(t), u_i(t), \omega_i(t), t\big), \\ y_i(t) &= h_i\big(x_i(t), v_i(t), t\big), \end{aligned} \tag{3}$$

where $x_i \in \mathbb{R}^{n_i}$ is the state vector, $u_i \in \mathbb{R}^{m_i}$ is the control input, $y_i \in \mathbb{R}^{p_i}$ is the measured output, and ω_i, v_i denote process and measurement disturbances, respectively. In the cooperative control setting, the control protocol u_i is typically a function of local relative information, i.e., $u_i = \kappa_i(\Sigma_{j \in \mathcal{N}_i} a_{ij}(x_j - x_i))$, where κ_i is a feedback law to be designed.

Depending on the physical characteristics of the agent, the functions f_i and h_i can be concretized into several widely studied categories. Agricultural embodied intelligent robot swarms are typically heterogeneous, consisting of nonholonomic ground vehicles for precise spraying and weeding, as well as underactuated aerial platforms for crop monitoring and mapping. Ground robots are often modeled as EL systems with nonholonomic constraints, while aerial robots exhibit strong nonlinearity and underactuated dynamics. This heterogeneity brings additional complexity to resilient consensus and formation control: different platforms have different driving capabilities, state dimensions, and vulnerability characteristics, thereby expanding the attack surface and increasing the difficulty of designing a unified resilient protocol.

To capture the heterogeneous collaboration between agricultural machinery and UAV formations, we model the ground–air swarm as a switched heterogeneous MAS. Ground vehicles are described by EL equations with nonholonomic constraints, while UAVs are modeled by underactuated nonlinear dynamics. The interconnection between ground and aerial agents is not merely informational but also physical: UAVs provide overhead perception and communication relaying, while ground robots execute physical tasks such as spraying, weeding, and harvesting. This physical–informational coupling creates new attack surfaces. For example, a compromised UAV can inject false crop-health maps into the ground robot network, causing incorrect spraying decisions; conversely, a compromised ground robot can broadcast false position or task-completion information, misleading UAV scheduling. Secure control must therefore consider not only consensus and formation stability but also task-level semantic consistency and physical safety constraints such as sprayer pressure limits, weeder torque bounds, and obstacle avoidance margins.

The simplest and most extensively studied case is that of homogeneous linear time-invariant (LTI) agents[66]:

$$\dot{x}_i(t) = Ax_i(t) + Bu_i(t), \qquad i = 1, \ldots, N, \tag{4}$$

where $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ are constant matrices. When agents possess different physical characteristics, the MAS is said to be heterogeneous, and the dynamics take the form

$$\dot{x}_i(t) = A_i x_i(t) + B_i u_i(t), \tag{5}$$

with $A_i \in \mathbb{R}^{n_i\times n_i}$, $B_i \in \mathbb{R}^{n_i\times m_i}$. Heterogeneity introduces additional complexity because the agents' state dimensions and dynamics may differ; output regulation and consensus require appropriate internal model or observer-based compensators[66,68].

A broad class of nonlinear agents encountered in robotic and aerospace applications can be modeled as affine-in-control nonlinear systems:

$$\dot{x}_i = f_i(x_i) + g_i(x_i)u_i, \tag{6}$$

where f_i: $\mathbb{R}^{n_i} \to \mathbb{R}^{n_i}$ and g_i: $\mathbb{R}^{n_i} \to \mathbb{R}^{n_i\times m_i}$ are locally Lipschitz. This structure enables feedback linearization, SMC, and adaptive designs[69,70].

A particularly important subclass for backstepping-based control is the strict-feedback form:

$$\begin{aligned} \dot{x}_{i,1} &= x_{i,2} + \varphi_{i,1}^{\top}(x_{i,1})\theta_i, \\ \dot{x}_{i,2} &= x_{i,3} + \varphi_{i,2}^{\top}(x_{i,1}, x_{i,2})\theta_i, \\ &\vdots \\ \dot{x}_{i,n_i} &= b_i u_i + \varphi_{i,n_i}^{\top}(x_i)\theta_i, \end{aligned} \tag{7}$$

where $\theta_i \in \mathbb{R}^{q_i}$ is a vector of unknown constant parameters, $\varphi_{i,k}$ are known smooth function vectors, and $b_i \neq 0$ is a known constant gain. This model is widely used for adaptive fuzzy and NN control of uncertain nonlinear MASs under cyberattacks[36,37].

Robotic manipulators, autonomous vehicles, and spacecraft are typically governed by EL equations of the form

$$M_i(q_i)\ddot{q}_i + C_i(q_i, \dot{q}_i)\dot{q}_i + g_i(q_i) = \tau_i + \tau_{d,i}, \tag{8}$$

where $q_i \in \mathbb{R}^{n_i}$ is the vector of generalized coordinates, $M_i(q_i)$ is the positive definite inertia matrix, $C_i(q_i, \dot{q}_i)\dot{q}_i$ represents Coriolis and centrifugal forces, $g_i(q_i)$ is the gravitational torque, τ_i is the control input torque, and $\tau_{d,i}$ is an external disturbance. The EL model possesses several well-known structural properties, namely, the skew-symmetry of $\dot{M}_i - 2C_i$ and linearity in a set of physical parameters, that are essential for adaptive and passivity-based control design[71,72].

In recent years, fractional-order calculus has been employed to model the memory and hereditary properties of certain physical processes, such as diffusion waves and viscoelasticity. The fractional-order MAS dynamics are described by the Caputo fractional derivative:

$$ {}_0^C D_t^{\alpha} x_i(t) = f_i(x_i(t), u_i(t), t), \qquad 0 < \alpha < 2, \tag{9}$$

where ${}_0{}^C D_t^{\alpha}$ denotes the Caputo fractional derivative of order α[73]. When $\alpha = 1$, the system reduces to the standard integer-order model. The non-local nature of fractional operators poses significant challenges for stability analysis under DoS attacks, requiring extended Lyapunov methods and memory-fusion controllers.

Communication noise and random cyberattacks naturally introduce stochastic perturbations. A general stochastic nonlinear MAS can be described by a set of Itô stochastic differential equations:

$$dx_i(t) = f_i(x_i(t), u_i(t))dt + g_i(x_i(t))dW_i(t), \tag{10}$$

where $W_i(t)$ denotes a standard Wiener process independent across agents, and the functions f_i, g_i satisfy appropriate Lipschitz and growth conditions to ensure existence and uniqueness of solutions[74,75]. Stochastic stability notions, such as noise-to-state stability and mean-square consensus, are then employed to evaluate system performance under random DoS attacks.

Practical agents often face actuator imperfections that alter the intended control input. A commonly adopted actuator fault model is given by

$$u_i^F(t) = \rho_i(t)u_i(t) + \phi_i(t), \tag{11}$$

where $\rho_i(t) \in [\underline{\rho}_i, 1]$ is the multiplicative efficiency factor ($0 < \underline{\rho}_i \leq 1$) and $\phi_i(t)$ is an additive stuck/bias fault[76,77]. When $\rho_i = 1$ and $\phi_i \neq 0$, the actuator is under bias fault; when $\rho_i = 0$, the actuator loses effectiveness completely. In addition, the physical limits of actuators impose amplitude saturation:

$$\mathrm{sat}(u_i) = \mathrm{sign}(u_i)\min\{|u_i|, \bar{u}_i\}, \tag{12}$$

where $\bar{u}_i > 0$ is the maximum admissible control magnitude[78,79]. Both faults and saturation strongly interact with security mechanisms because DoS or deception attacks may prevent the timely diagnosis and accommodation of such hardware anomalies.

2.3. Cyber attack models: a unified taxonomy

Traditional IT security primarily protects the confidentiality, integrity, and availability of information through cryptographic primitives, authentication, firewalls, and access control. In contrast, control-oriented security for MASs must additionally guarantee the stability, performance, and physical safety of the closed-loop dynamics under adversarial behavior[80]. The essential difference is that control-oriented security treats attacks as signals injected into the feedback loop, and therefore models, detects, and compensates for them using control-theoretic tools such as observers, robust control, adaptive laws, and resilient protocols. This perspective exploits system dynamics, redundancy, and physical invariants that are largely absent in traditional IT security.

Figure 2 illustrates the complementary roles of the two paradigms within a feedback loop. Traditional IT security mainly operates at the transmission and computing infrastructure, protecting data availability, integrity, and confidentiality. Control-oriented security is distributed across the four information-flow nodes - perception, transmission, decision-making, and execution - and is tightly coupled with the physical plant. For example, SSE and physical-layer trust act at the perception node; event-triggered, self-triggered, and quantized communication act at the transmission node; resilient consensus, observer-based control, adaptive control, and RL act at the decision node; and fault-tolerant control, SMC, Nussbaum-function compensation, and impulsive control act at the execution node. Cross-layer mechanisms, such as watermarking, digital twins, and moving target defense, bridge multiple nodes. This diagram helps readers in the cybersecurity community see that control-oriented security is not a replacement for information technology (IT) security but a complementary layer that protects against the physical consequences of cyberattacks.

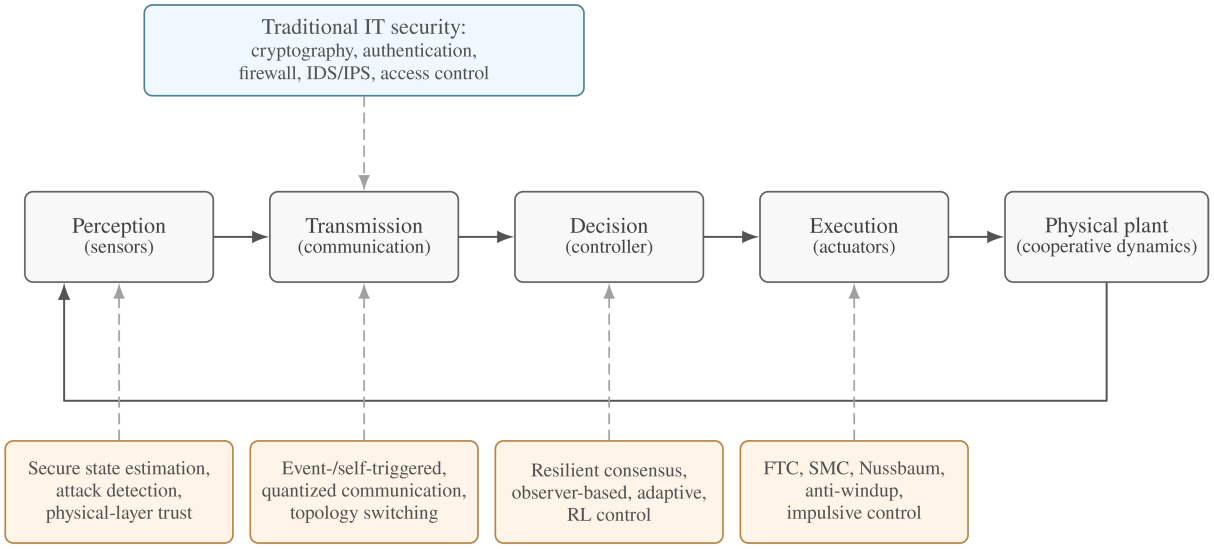


Figure 2. Complementary roles of traditional IT security and control-oriented security in a feedback loop. Traditional IT security mainly protects the communication and computing infrastructure. Control-oriented security is distributed across perception, transmission, decision-making, and execution, and exploits physical dynamics, observers, and resilient controllers to preserve stability, performance, and safety. IT: Information technology; IDS/IPS: intrusion detection system/intrusion prevention system; RL: reinforcement learning; FTC: fault-tolerant control; SMC: sliding mode control.

2.3.1. A unified information-flow attack framework

Recall from Section 2.2 that the closed-loop operation of an MAS can be abstracted as a sequential information flow: (i) perception (sensors), (ii) transmission (communication network), (iii) decision (controller/estimator), and (iv) execution (actuators). Let $\mathcal{I}_P$, $\mathcal{I}_T$, $\mathcal{I}_D$, and $\mathcal{I}_E$ denote the sets of information-bearing signals at each stage, respectively. A cyberattack on the MAS is modeled as a bounded-energy operator $\mathcal{A}$ that acts on a subset of these signals, producing a corrupted version[23]:

$$\tilde{s}(t) = \mathcal{A}\big(s(t), \theta_a(t), t\big), \tag{13}$$

where $s \in \mathcal{I}_*$, $\tilde{s}$ is the attacked signal, and $\theta_a(t) \in \Theta$ denotes the attack parameters, which may include timing, injected data, and energy constraints. This abstraction unifies diverse attack types: a DoS attack on the transmission stage sets $\tilde{s} = \emptyset$ (erasure), a deception attack on perception sets $\tilde{y} = y + \delta$, and a Byzantine attack on decision compromises the control logic itself.

The proposed unified attack framework is a modeling language for locating attacks within the closed-loop information flow, rather than a claim that DoS, FDI, replay, and Byzantine attacks share the same mathematical characteristics. These attack classes differ substantially in their action sets, state dependence, stealthiness, and detectability, and they require different mathematical tools for analysis and defense design.

Specifically, a DoS attack is a set-valued erasure or blocking operator: on an attacked link, the received signal is either erased ($\tilde{y}_{ij} = \emptyset$) or held at its previous value, and the attack is characterized by switching instants, frequency, and duration[23,24]. An FDI attack is an additive or multiplicative injection operator: $\tilde{y}_{ij} = y_j + \delta_{ij}^a$ or $\tilde{u}_i = \rho_i^a u_i + \phi_i^a$, where the attack signal may be state-dependent and designed to lie within the system's dynamic range to evade residual-based detectors. A replay attack is a time-delay substitution operator: $\tilde{y}_{ij}(t) = y_j(t - T_R)$, which preserves signal consistency with historical behavior and is therefore stealthy under static detectors. A Byzantine attack is an arbitrary behavior operator: a compromised agent may send inconsistent

information to different neighbors, and its output is not constrained by any nominal dynamics, making it the most general and least structured attack class.

The unified operator $\mathcal{A}(\cdot)$ therefore serves three purposes. First, it provides a common notation to specify which information-bearing signal at which node is corrupted, namely perception, transmission, decision-making, or execution. Second, it enables a systematic comparison of defense mechanisms in terms of the attack assumptions they require, the mathematical tools they employ, and the stability or security guarantees they provide. Third, it exposes the security–performance–resource trade-offs associated with each attack class. For example, DoS resilience is typically analyzed via switched systems and average dwell-time (ADT) conditions; FDI resilience relies on observers, robust control, and watermarking; replay resilience requires time-stamping, challenge-response, or physical-layer fingerprints; and Byzantine resilience is built on redundancy, median-type consensus, and graph-theoretic conditions such as *F*-local or *F*-total models.

In other words, the unified framework does not imply that these attacks can be handled by identical mathematical tools. Rather, it provides a common coordinate system that maps each attack to the node where it acts and each defense to the node where it restores availability, integrity, or physical safety. This mapping is what enables the cross-layer coupling analysis and the resource-aware defense taxonomy developed in the remainder of this survey. We therefore treat DoS, FDI, replay, Byzantine, actuator, and hybrid attacks with their respective mathematical models, while using the four-node architecture as the unifying organizational lens.

2.3.2. DoS attacks

DoS attacks aim to block information transmission by jamming communication channels, flooding the network with superfluous requests, or exploiting protocol vulnerabilities. At the transmission layer, a DoS attack on the link from agent j to agent i is characterized by a time-dependent attack indicator $\theta_{ij}(t) \in \{0, 1\}$[24]:

$$\tilde{y}_{ij}(t) = \begin{cases} y_j(t), & \text{if } \theta_{ij}(t) = 1 \text{ (communication active)}, \\ \emptyset, & \text{if } \theta_{ij}(t) = 0 \text{ (under attack)}, \end{cases} \tag{14}$$

where $\tilde{y}_{ij}$ is the information actually received by agent i from neighbor j. The attack signal θ_{ij} is determined by the adversary's energy budget and strategy.

The most widely adopted deterministic model, following the seminal work of De Persis and Tesi[24], imposes explicit bounds on the attack frequency and duration. Let $\{t_k\}_{k\in\mathbb{N}}$ be the time instants at which the attack switches between on ($\theta_{ij} = 0$) and off ($\theta_{ij} = 1$) states. Define the sequence of DoS off-intervals as $\mathcal{I}_k = [t_k, t_k + \tau_k)$, during which communication is blocked. The attack is said to satisfy the frequency–duration constraints if there exist constants $\kappa \geq 0$ and $\tau_D > 0$ such that

$$n(t, s) \leq \kappa + \frac{t - s}{\tau_D}, \qquad \forall\, t \geq s \geq 0, \tag{15}$$

where $n(t, s)$ denotes the number of DoS off-intervals in $[s, t)$, and

$$\sum_{k:\mathcal{I}_k \subset [s,t)} |\mathcal{I}_k| \leq \Xi_D + \eta(t - s), \tag{16}$$

with $\eta \in (0, 1)$ being the attack duty cycle and $\Xi_D \geq 0$ a constant[24,81]. These constraints guarantee that communication blackouts cannot be too frequent or too long, enabling robust consensus design via switched system theory.

When attack patterns are unpredictable or lack deterministic bounds, stochastic models are employed. A common approach is to model the attack indicator on each link as a Bernoulli random process:

$$\theta_{ij}(t) \sim \text{Bernoulli}(p_{ij}), \tag{17}$$

where $p_{ij} \in [0, 1]$ is the probability of a successful transmission[52]. More sophisticated models treat the network-wide attack state as a continuous-time Markov chain with a finite number of modes, each corresponding to a subset of blocked links. In this setting, consensus is analyzed in the mean-square or almost-sure sense.

In large-scale MASs, adversaries may launch independent DoS attacks on different communication channels asynchronously, i.e., $\theta_{ij}(t)$ and $\theta_{kl}(t)$ are not synchronized. The overall attacked topology $\mathcal{G}(t) = (\mathcal{V}, \mathcal{E}(t))$ then evolves over time, and the union graph condition (see Section 2.1) becomes the key to preserving connectivity[27,28].

DoS attacks differ fundamentally from traditional packet loss. Traditional packet loss is usually modeled as random and independent events with known probability distributions (such as the Bernoulli or Gilbert-Elliott models), mainly resulting from channel fading, interference, or congestion. In contrast, DoS attacks are intentional and energy-constrained, often showing temporal correlation and strategic patterns, such as sudden disruptions synchronized with control updates[12,49]. Attackers can adaptively adjust interference timing to maximize destructive effect, while natural packet loss is smooth and independent of the control protocol. This distinction prompts the adoption of frequency-duration constraints, random game models, and adversarial scheduling assumptions in anti-DoS designs.

2.3.3. FDI and replay attacks

FDI attacks compromise data integrity without necessarily disrupting communication availability. The adversary injects false data, replays previous signals, or entirely replaces transmitted measurements to mislead neighboring agents. In an FDI attack, the transmitted signal y_j from agent j is corrupted additively:

$$\tilde{y}_{ij}(t) = y_j(t) + \delta_{ij}^a(t), \tag{18}$$

where $\delta_{ij}{}^a(t)$ is the attack injection signal designed by the adversary[53,82]. When the attack acts directly on the control input or the state estimate, it may take a multiplicative form:

$$\tilde{u}_i(t) = \rho_i^a(t)u_i(t) + \phi_i^a(t), \tag{19}$$

with $\rho_i{}^a(t) \neq 1$ or $\phi_i{}^a(t) \neq 0$.

In a replay attack, the adversary records previous measurement streams and replays them at a later time:

$$\tilde{y}_{ij}(t) = y_j(t - T_R), \tag{20}$$

where $T_R > 0$ is the replay delay. Replay attacks are particularly stealthy because they inject signals consistent with the system's historical behavior[83]. A more aggressive deception attack replaces the true signal entirely with an adversary-designed sequence $a_{ij}(t)$:

$$\tilde{y}_{ij}(t) = a_{ij}(t), \tag{21}$$

which can be modeled as Byzantine behavior on the communication channel.

FDI attacks differ fundamentally from traditional noise and disturbances. Traditional noise and disturbances are usually bounded or random processes with known statistical characteristics, and are independent of the system state. In contrast, FDI attacks are state-dependent, covert, and aim to exploit system knowledge to circumvent detectors[23]. Attackers can inject signals within the system's dynamic value range, making them difficult to distinguish from legitimate control inputs under traditional residual tests. This feature has driven the development of SSE, watermarking technology, and physical-layer trust mechanisms, which utilize redundant information and physical invariants to detect and isolate such malicious injections.

2.3.4. Byzantine (malicious node) attacks

Instead of attacking communication links, an adversary may compromise entire agents, turning them into Byzantine nodes that send arbitrary and potentially inconsistent information to different neighbors[84]. Formally, let $\mathcal{F} \subset \mathcal{V}$ be the set of malicious nodes. The output of a Byzantine agent $j \in \mathcal{F}$ is no longer governed by its nominal dynamics but is replaced by an adversarial signal $\tilde{y}_j(t)$:

$$\tilde{y}_j(t) = \mathcal{B}_j\left(y_j(t), t, \{\tilde{y}_k\}_{k\in\mathcal{F}\setminus\{j\}}\right), \tag{22}$$

where $\mathcal{B}_j$ is an arbitrary (potentially time-varying) function chosen by the adversary.

Two standard models quantify the adversarial power[14,85]: (i) the F-local model, in which for every normal agent i, at most F of its neighbors are malicious, i.e., $|\mathcal{N}_i \cap \mathcal{F}| \le F$; and (ii) the F-total model, in which the total number of malicious agents in the entire network satisfies $|\mathcal{F}| \le F$. Resilient consensus algorithms, such as the mean-subsequence-reduced (MSR) family, are typically designed to tolerate up to F malicious agents under these models[14].

2.3.5. Actuator attacks and hybrid composite attacks

Actuator attacks target the execution layer and directly affect physical actuation. They are modeled as multiplicative or additive faults, possibly combined with saturation, as in Equations (11) and (12). The adversary may have physical access or may inject false commands through compromised software. Unlike communication-layer attacks, actuator attacks have immediate physical consequences and require safety-critical guarantees.

In realistic scenarios, multiple attack types may coexist. For example, a DoS attack can be launched on some links to degrade connectivity, while simultaneously FDI attacks corrupt the data on the remaining links[56,57]. Such hybrid attacks are modeled by combining the corresponding operators:

$$\tilde{y}_{ij}(t) = \theta_{ij}(t)\left[y_j(t) + \delta_{ij}^a(t)\right]. \tag{23}$$

When DoS is active ($\theta_{ij} = 0$), the deception injection becomes irrelevant; when DoS is inactive, the receiver sees a corrupted signal. More generally, sensor and actuator attacks may occur independently of communication-layer attacks, requiring multi-layer defense frameworks[21,56].

2.3.6. Agricultural-specific attack models

For agricultural robot swarms[10], the attack model must be specialized beyond generic DoS, FDI, replay, and Byzantine categories. GNSS spoofing is a perception-layer deception attack that can cause seeding overlap, missed crop rows, or incorrect sprayer activation[44,53]. Multispectral deception injects false crop-health

Table 2. Unified taxonomy of cyber attacks in MASs

Attack type	Target layer	Compromised property	Mathematical model	Key parameters	Representative references
DoS (periodic and aperiodic)	Transmission	Availability	$\theta_{ij}(t) \in \{0, 1\}$	Frequency, duration, duty cycle	[24,81]
Stochastic DoS	Transmission	Availability	Bernoulli/Markov chain	Probability p_{ij}	[52]
Asynchronous multi-channel DoS	Transmission	Availability	Independent $\theta_{ij}(t)$ per link	Per-link switching	[27,28]
FDI (additive/multiplicative)	Transmission/perception	Integrity	$\tilde{y} = y + \delta^a$; $\tilde{u} = \rho^a u + \phi^a$	Injection δ^a, ρ^a, ϕ^a	[53,82]
Replay attack	Transmission/perception	Integrity	$\tilde{y}(t) = y(t - T_R)$	Delay T_R	[83]
Byzantine (F-local)	Decision	Integrity/behavior	$\tilde{y}_j = \mathcal{B}_j(\cdot)$	F (max. per-agent)	[85]
Byzantine (F-total)	Decision	Integrity/behavior	$\tilde{y}_j = \mathcal{B}_j(\cdot)$	F (total)	[14]
Actuator attack	Execution	Integrity/behavior	$u^F = \rho u + \phi$, sat(u)	ρ, ϕ, $\bar{u}$	[76,77]
Hybrid DoS + FDI	Transmission	Availability and integrity	$\tilde{y} = \theta(y + \delta^a)$	All above	[56,57]

MASs: Multi-agent systems; DoS: denial-of-service; FDI: false data injection.

indices, leading to unnecessary irrigation or pesticide application. Camera spoofing imitates crop rows or obstacles, causing navigation errors. LiDAR and IMU replay mislead simultaneous localization and mapping (SLAM) in orchards and greenhouses where GNSS is unreliable[83]. Actuator attacks on sprayer pumps, weeders, and wheel motors have immediate physical consequences, including chemical over-application, crop damage, and uneven field coverage[76]. Communication-layer attacks exploit intermittent rural connectivity, long-range fading, and canopy obstruction to disrupt coordination[27]. These agricultural-specific attack models are mapped to the four-node information-flow architecture[46,62]: perception-layer deception, transmission-layer jamming and packet loss, decision-layer Byzantine or FDI-compromised neighbors, and execution-layer actuator attacks. This mapping ensures that the application perspective is integrated into the technical core of the survey rather than treated as a separate example.

2.3.7. Attack taxonomy summary

Table 2 presents a compact overview of the attack types discussed above, classified by operational level, compromised property (availability, integrity, or both), and the mathematical model employed throughout this survey. The table highlights a fundamental distinction: DoS and asynchronous multi-channel attacks compromise availability through set-valued operators, whereas FDI, replay, and Byzantine attacks compromise integrity through injection, substitution, or arbitrary behavior operators. Actuator and hybrid attacks cut across both categories and require cross-layer defenses.

2.4. Control objectives and the four-node information flow architecture

Having established the graph-theoretic foundations, agent dynamics, and a unified taxonomy of cyberattacks, we now formalize the cooperative control objectives pursued by MASs and introduce the four-node information flow architecture that will serve as the backbone for our defense strategy classification.

Let the agents be indexed by $i = 1, 2, \ldots, N$. Depending on the mission, the MAS is required to achieve one or more of the following objectives through purely local information exchange.

Definition 2[50,51]**.** The system is said to achieve state consensus if for any initial conditions,

$$\lim_{t\to\infty} \|x_i(t) - x_j(t)\| = 0, \quad \forall i, j \in \mathcal{V}. \tag{24}$$

In the leader–follower setting, leader-follower consensus requires $\lim_{t\to\infty}||x_i(t) - x_0(t)|| = 0$ for all followers i, where $x_0(t)$ is the leader's state.

Definition 3[86,87]**.** The MAS achieves a desired time-invariant formation specified by relative offset vectors $d_{ij} \in \mathbb{R}^n$ if

$$\lim_{t\to\infty} \left(x_i(t) - x_j(t)\right) = d_{ij}, \quad \forall i, j \in \mathcal{V}. \tag{25}$$

More general time-varying and affine formations have also been studied.

Definition 4[88,89]**.** Suppose the MAS consists of M followers and $N - M$ leaders. Let $\Omega_L(t) = \text{Co}\{x_{M+1}(t), \ldots, x_N(t)\}$ be the convex hull spanned by the leaders' states. Containment control is achieved if the followers' trajectories converge to the convex hull, i.e.,

$$\lim_{t\to\infty} \text{dist}\left(x_i(t), \Omega_L(t)\right) = 0, \quad i = 1, \ldots, M. \tag{26}$$

Definition 5[68,90,91]**.** Consider a linear exosystem generating the reference and disturbance signals: $\dot{w} = Sw$. Each agent's regulated output is $e_i = y_i - Qw$. The cooperative output regulation problem is to design distributed controllers such that, for all initial conditions,

$$\lim_{t\to\infty} e_i(t) = 0, \quad \forall i \in \mathcal{V}, \tag{27}$$

while maintaining internal stability.

Definition 6[92,93]**.** Let f_i: $\mathbb{R}^n \to \mathbb{R}$ be a local convex cost function known only to agent i. The objective of distributed optimization is to design local update laws such that the agents' states converge to a common minimizer of the global cost:

$$\lim_{t\to\infty} \|x_i(t) - x^*\| = 0, \quad x^* \in \arg\min_{x\in\mathbb{R}^n} \sum_{i=1}^{N} f_i(x). \tag{28}$$

Constraints and coupled costs can be incorporated.

In a cooperative control loop, information flows along a chain of four logical nodes: Perception (P), Transmission (T), Decision (D), and Execution (E). Figure 3 depicts this architecture for a single agent i interacting with its neighbors.

The signals at each node are defined as follows[20,94]:

• Perception node (P): The agent measures its own state x_i and receives transmitted information from neighbors through the communication network, denoted by $\hat{x}_{ji}$ for $j \in \mathcal{N}_i$. In an attack-free environment, $\hat{x}_{ji} = x_j$[44,95,96].

• Transmission node (T): The agent broadcasts its own state x_i over wireless links. The actual signal received by a neighbor j is $\tilde{x}_{ij}$, which may be corrupted by DoS or deception attacks as modeled in Section 2.3[63,97-100].

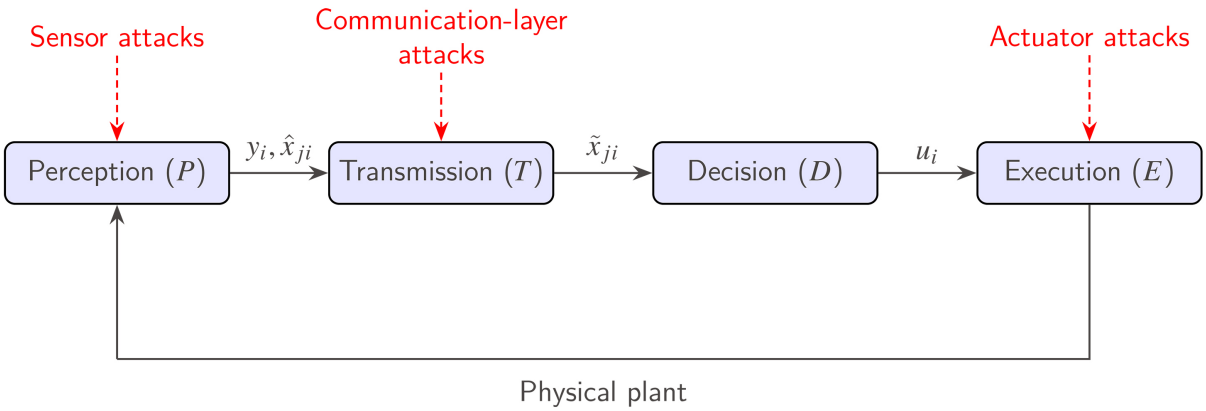


Figure 3. The four-node information-flow architecture of a cooperative control loop under cyberattacks. Red dashed lines indicate potential adversarial intrusion points.

Table 3. Defense strategy taxonomy based on the four-node information flow

Node	Defense objective	Methodologies	Attack addressed
Perception (P)	SSE; sensor attack detection	Observer-based estimation[111], attack detection filters[53], physical-layer trust, redundant sensors	Sensor deception, FDI on local measurements
Transmission (T)	Maintain communication availability and integrity despite attacks	Event-triggered scheduling[24], switching topologies[112], encryption, game-theoretic power control[113], physical-layer security	DoS, FDI on links, replay attacks, asynchronous multi-channel attacks
Decision (D)	Ensure correct coordination despite corrupted information	Resilient consensus (MSR, W-MSR)[14], robust optimization, adaptive neural/fuzzy control[36], RL-based defense	Byzantine nodes, FDI on received states, hybrid attacks
Execution (E)	Compensate for actuator attacks/faults	Fault-tolerant control[114], SMC[115], adaptive compensation, Nussbaum functions[116], robust saturation handling	Actuator FDI, multiplicative and additive faults

SSE: Secure state estimation; FDI: false data injection; DoS: denial-of-service; MSR: mean-subsequence-reduced; RL: reinforcement learning; SMC: sliding mode control.

• Decision node (D): The agent collects all available information, local measurement $\tilde{x}_i$ (possibly attacked at the sensor) and received neighbor states $\{\tilde{x}_{ji}\}$, and computes the control law $u_i = \kappa_i(\tilde{x}_i, \{\tilde{x}_{ji}\})$[101-105].

• Execution node (E): The control signal u_i is applied to the physical actuators. Actuator attacks, faults, or saturation may modify the effective input to $\tilde{u}_i$[106-110].

The four-node architecture provides a natural basis for classifying defense strategies by information-flow links. Table 3 presents the defense targets corresponding to each node, the representative methods reviewed in this survey, and the types of attacks they can resist. This classification system reveals complementary functional relationships among defense mechanisms and highlights the need to embed security from the very beginning of the design.

It is essential to distinguish single-layer defenses from cross-layer defenses. A single-layer defense acts primarily on one information-flow node, whereas a cross-layer defense couples two or more nodes and exploits their interactions. For example, event-triggered scheduling is a single-layer transmission defense; SSE is a single-layer perception defense; resilient consensus is a single-layer decision defense; and fault-tolerant control is a single-layer execution defense. In contrast, an event-triggered observer-based resilient controller is a cross-layer defense that couples transmission and decision-making; a digital twin-assisted

defense couples perception, decision-making, and execution; physical-layer trust couples perception and transmission; and integrated fault diagnosis with task reallocation couples execution and decision-making. The subsequent sections explicitly indicate the coupling structure of each defense family.

2.5. Cross-layer coupling and the security–performance–resource trade-off

Although the security–performance–resource trilemma is repeatedly emphasized throughout this survey, its manifestation in cross-layer joint defense mechanisms deserves a more systematic treatment. Most existing works optimize a single layer in isolation: perception-layer SSE improves estimation accuracy at the cost of additional computation and redundant sensing; transmission-layer event-triggered or quantized communication reduces bandwidth and exposure but may degrade the freshness and accuracy of information available to the decision layer; decision-layer learning-based or adaptive control enhances resilience to unknown attacks but requires substantial data, computation, and sometimes privacy-sensitive information exchange; execution-layer fault-tolerant or impulsive control preserves actuation safety but may increase control effort or reduce nominal performance. These single-layer optimizations are coupled through the closed-loop information flow and cannot be designed independently without incurring suboptimal or even unstable overall behavior.

We therefore summarize the intrinsic coupling effects among perception, transmission, decision-making, and execution as follows.

• Perception–Transmission coupling. The number and type of sensors determine the amount of information to be transmitted. Redundant or multimodal sensing improves attack detection but increases communication load and energy consumption. Conversely, aggressive event-triggered or quantized transmission reduces communication burden but may starve the perception layer of the data needed for SSE.

• Transmission–Decision coupling. DoS-induced packet loss and FDI-induced data corruption directly affect the information available to the decision layer. Event-triggered communication can reduce the attack surface, but the resulting intermittent information may degrade consensus accuracy and stability margins. Adaptive or learning-based decision layers may compensate for missing data, but they require additional computation and may suffer from parameter drift during long attack intervals.

• Decision–Execution coupling. Resilient decision policies often require greater control effort or more frequent switching, which may saturate actuators or accelerate hardware wear. Conversely, execution-layer faults or actuator attacks alter the effective control input, which the decision layer must estimate and compensate for. If the decision layer is not aware of execution-layer constraints, the resulting control may be infeasible or unsafe.

• Execution–Perception coupling. Actuator attacks or faults change the physical state of the plant, which in turn affects sensor measurements. Without cross-layer awareness, the perception layer may misinterpret attack-induced state deviations as sensor faults, leading to false alarms or missed detection.

A representative example of cross-layer joint defense is the hierarchical federated learning framework for anomaly detection in Industrial Internet of Things[62]. In that work, anomaly detection is performed at the perception layer, federated learning is used at the transmission layer to build a universal detection model without sharing raw local data, and deep RL is employed at the decision layer to optimize the integration weights of local models. The framework explicitly balances privacy leakage, detection accuracy, system throughput, and latency. This illustrates that cross-layer coupling is not merely a theoretical concern but a practical design requirement: improving detection accuracy at the perception layer may increase

Table 4. Cross-layer coupling effects and trade-offs among perception, transmission, decision-making, and execution

Coupling	Primary interaction	Security benefit	Resource cost	Potential performance degradation
Perception–transmission	Sensor redundancy *vs.* communication load[99]	Improved attack detection and state estimation	Higher bandwidth, energy, and computation	Increased latency, congestion, exposure to DoS
Transmission–decision	Information freshness and integrity *vs.* control performance[31]	Reduced attack surface, resilient consensus	Computation for compensation and learning	Degraded consensus accuracy, parameter drift, slower convergence
Decision–execution	Control effort *vs.* actuator constraints[37]	Fault-tolerant and attack-resilient actuation	Actuator wear, saturation risk, higher control energy	Reduced nominal performance, chattering, saturation
Execution–perception	Physical state evolution *vs.* sensor interpretation[46]	Detection of stealthy actuator attacks	Additional sensing and diagnosis	False alarms, misclassification of faults as attacks
Cross-layer joint defense	Coordinated optimization across nodes[62]	Balanced resilience, privacy, and efficiency	Joint computation and communication overhead	Design complexity, need for formal guarantees

DoS: Denial-of-service.

communication and computation overhead at the transmission and decision layers, while reducing communication overhead may degrade detection accuracy and resilience.

Table 4 summarizes the coupling effects among the four nodes, the associated resource costs, the security benefits, and the potential performance degradation.

2.6. Relation to existing CPS/MAS security taxonomies

To further justify the novelty of the four-node information-flow architecture, we explicitly compare it with representative existing taxonomies in cyber-physical systems and secure control.

Surveys such as[46-48] classify attacks into DoS, deception, replay, and disclosure categories, or distinguish between integrity, availability, and confidentiality attacks. These taxonomies are valuable for characterizing adversary capabilities and mathematical models. However, they do not indicate where in the control loop a defense should be deployed, nor how a defense at one layer affects another. For example, an event-triggered, observer-based controller acts simultaneously at the transmission node by reducing exposure and at the decision node by reconstructing states; this cross-layer coupling is invisible in an attack-type taxonomy.

Recent work has proposed architecture-driven taxonomies such as the ADHL framework[60], which classifies threats and defenses across device, network, control, data, algorithm, and application layers, and the layered attack surface model (LASM)[61], which decomposes agentic AI systems into seven layers. These frameworks provide valuable architectural decomposition, but they are primarily technology-stack-oriented: they describe the layers of a computing or robotic system rather than the functional stages of a closed-loop cooperative control task. In contrast, the four-node architecture is loop-centric: perception, transmission, decision-making, and execution correspond to the stages through which information flows to generate a control action, and the coupling among these stages is explicitly modeled. This distinction is particularly important for MASs, where the same physical node may participate in multiple loops, and the same attack may propagate across nodes.

Other surveys organize defenses by methodology, such as detection-based, estimation-based, robust-control-based, and learning-based approaches[21]. While useful for methodological comparison, such taxonomies do not reveal the functional role of each defense within the information flow. A detection filter and a resilient consensus protocol may both be “detection-based” or “estimation-based”, yet they act at different nodes and

have different coupling structures. The four-node architecture clarifies this by mapping each defense to the node where it operates and by distinguishing single-layer from cross-layer mechanisms.

The comparison above highlights three advantages that cannot be obtained from existing classifications. First, the four-node architecture is explicitly feedback-loop-centric, which enables the analysis of cross-layer coupling effects such as perception–transmission trade-offs in multimodal sensing and transmission–decision trade-offs in event-triggered resilient control[31,62]. Second, it provides a common coordinate system for comparing heterogeneous defense mechanisms - observer-based, event-triggered, adaptive, learning-based, and fault-tolerant - on the same functional basis, namely the node at which they restore availability, integrity, or physical safety[14,21,46,114]. Third, it exposes the security–performance–resource trilemma at each node and across nodes, which is essential for resource-aware co-design in resource-constrained platforms such as UAVs and agricultural robot swarms[10,99]. These insights are not directly available from attack-type, architecture-layer, or defense-tool taxonomies[47,60,61], and they motivate the cross-layer coupling analysis and the resource-aware defense taxonomy developed in Sections 3 and 4.

3. RESILIENT DEFENSE MECHANISMS ALONG THE INFORMATION FLOW

This section reviews resilient defense mechanisms according to the four-node information-flow architecture. For each node, we first define the defense objective, then summarize representative single-layer methods, and finally discuss cross-layer mechanisms that couple this node with others. To avoid a purely descriptive presentation, each defense family is compared along six dimensions: (i) assumptions on system dynamics and adversary knowledge; (ii) the attack model it addresses; (iii) the control objective; (iv) the stability, convergence, or security guarantee; (v) communication and computation requirements; and (vi) the validation method, ranging from numerical simulation to hardware-in-the-loop testing and field trials. These dimensions are used consistently across Sections 3.1-3.5, and each summary table is accompanied by a critical comparison in the text, so that competing approaches can be compared on a common basis rather than merely enumerated.

3.1. Perception-layer defense: SSE under compromised sensors

This subsection reviews perception-layer defenses, emphasizing SSE, attack detection, and physical-layer trust. We compare centralized and distributed estimators in terms of observability requirements, sparsity assumptions, computational complexity, and resilience guarantees.

The perception layer constitutes the sensory interface through which each agent measures its own state, the environment, and potentially the states of neighboring agents. In an ideal setting, the measurement $y_i(t) \in \mathbb{R}^{p_i}$ of agent i is a deterministic function of its true state $x_i(t)$, e.g., $y_i(t) = C_i x_i(t)$. However, as shown in Figure 3, the perception node is a primary attack target for adversaries who seek to corrupt the information basis of cooperative control at its source. When sensor measurements are compromised, all downstream processing (including state estimation, control law computation, and coordination) becomes unreliable. The primary objective of perception-layer defense is therefore to obtain trustworthy state estimates despite compromised sensors, thereby insulating the decision and execution layers from corrupted data.

Sensor attacks on agent i can be modeled as alterations of the nominal measurement equation. Let $y_i(t) = h_i(x_i(t))$ be the true sensor output. Under attack, the actual measurement available to the agent becomes

$$\tilde{y}_i(t) = y_i(t) + \delta_i^s(t) \quad \text{or} \quad \tilde{y}_i(t) = a_i^s(t), \tag{29}$$

where $\delta_i^s(t)$ is an additive injection and $a_i^s(t)$ is an arbitrary replacement signal. The attack is sparse if only a subset $\mathcal{K} \subset \{1, \ldots, N\}$ of agents are affected, with $|\mathcal{K}| \leq \kappa$[111]. Such sparsity is a critical structural property

exploited by many secure estimation algorithms. A more general model considers s-sparse attacks on the vector of all sensor measurements: the attacker can corrupt at most s out of p sensors of a single agent or across the network[53,117]. Additionally, the attack may be stealthy, i.e., designed to evade residual-based detectors. A common stealthiness condition is that the attack sequence δ_i^s lies in the range space of a suitable system matrix, thus producing zero innovation under the nominal observer. Alternatively, the adversary may launch replay attacks by injecting previously recorded sensor traces[83]. These sophisticated attacks demand detection mechanisms beyond conventional static thresholds.

The fundamental problem of SSE is to reconstruct the true state $x_i(t)$ from a set of possibly corrupted measurements $\{\tilde{y}_i\}$, with guarantees of correctness provided that the number of attacked sensors does not exceed a certain threshold. Early SSE work focused on centralized cyber-physical systems and can be broadly classified into two categories. (i) Compressive sensing and ℓ_0/ℓ_1 optimization: by modeling the attack vector as a sparse signal, the SSE problem is formulated as

$$\min_{\hat{x},\hat{\delta}} \|\hat{\delta}\|_0 \quad \text{s.t.} \quad \tilde{y} = O\hat{x} + \hat{\delta}, \tag{30}$$

where O is the observability matrix. Under appropriate observability and redundancy conditions, the true state can be recovered exactly by solving a convex relaxation[111,118]; (ii) Satisfiability modulo theory (SMT) and set-based methods: SMT solvers search for a state consistent with all but s measurements, yielding a guaranteed resilient estimate whenever fewer than s sensors are attacked[117,119].

In MASs, the distributed nature of information adds an extra layer of complexity: each agent can only rely on its own sensor measurements and neighbor data that may have been tampered with during transmission. To address this challenge, two complementary research approaches have emerged: security estimation based on distributed observers and resilient consensus based on local estimation. A common design approach is to deploy a set of local observers, each of which estimates the state of a portion of the agent, and then integrate these estimation results through a resilient consensus mechanism. For instance, the sampled-data adaptive distributed observer proposed by Deng and Wen[29] utilizes buffers to accommodate non-uniform time delays and prevents attackers from understanding the sampling patterns through time-varying sampling sequences, thereby reconstructing the state of the leader in heterogeneous linear MASs under the condition where DoS and spoofing attacks coexist. To ensure the smoothness of the observer derivative (which is a key prerequisite for implementing backstepping control under DoS attacks), Zhang *et al.* constructed a resilient observer for uncertain nonlinear MASs with strict feedback forms, specifically implemented as a first-order low-pass filter with an upper triangular chain structure[102]. Another paradigm regards the local state estimation of each agent as a scalar or vector form of the quantity to be agreed upon, and uses resilient consensus algorithms to eliminate extreme or malicious estimates. The MSR algorithm and its weighted variant W-MSR guarantee that, as long as the network meets certain robustness conditions, estimates of normal agents converge to the convex hull of the initial values of themselves and other normal agents[14,85]. When these resilient estimators are further combined with event-triggered communication, the attack surface can be significantly reduced by restricting non-essential transmissions[120,121].

Shang[59] extends resilient consensus to mobile malicious attacks on random dynamic networks. The adversary can compromise different agents over time, and the network topology itself is random. The proposed distributed discrete-time algorithms achieve resilient vector consensus under stochastic robustness conditions, using Lyapunov theory and martingale convergence. Compared with MSR/W-MSR algorithms on static or deterministically switching graphs, this line of work is more general in both the adversary model (mobile rather than static) and the network model (random rather than deterministic). However, the guarantee is asymptotic in expectation, and the work does not address communication-layer availability

attacks, FDI on links, or execution-layer faults. Within the four-node framework, it is a single-layer decision defense. Integrating it with transmission-layer event-triggering and perception-layer secure estimation remains an open direction.

Perception-layer detection mechanisms aim to raise an alarm when sensor measurements deviate from expected behavior. Classical model-based methods rely on the generation of a residual signal $r_i(t) = \tilde{y}_i(t) - C_i\hat{x}_i(t)$, where $\hat{x}_i$ is the state estimate from a nominal observer. If the system is attack-free, the residual should behave as white noise or exhibit bounded fluctuations; a sudden statistical change indicates a potential attack. Anomaly detectors in secure MASs span a spectrum of design philosophies. The chi-square (χ^2) detector compares the weighted norm of the residual against a threshold derived from the χ^2 distribution, proving effective against non-stealthy attacks; however, adversaries with knowledge of the system matrices can completely bypass it. For more subtle, time-evolving threats, cumulative sum (CUSUM) and sequential probability ratio tests track persistent changes in the residual distribution over time, offering heightened sensitivity to slow-drift attacks that would otherwise remain unnoticed[122]. In contrast, set-based and interval observers depart from the paradigm of a single point estimate, instead maintaining the set of all states that are consistent with the model and measurements, and triggering an alarm when this guaranteed set becomes empty or unacceptably large, thereby providing a rigorous framework for resilience under bounded uncertainties[123].

In a multi-agent environment, cross-checking among neighbors can further improve the detection capability. For instance, an agent can compare a neighbor's broadcast state with its own prediction of the neighbor's trajectory. Persistent inconsistency indicates that either the agent's own sensors have been attacked or the communication channel itself has been attacked. In addition, machine learning classifiers such as support vector machines and deep autoencoders have been used to learn normal behavior patterns in sensor data streams, marking anomalies without explicit system models.

In addition to pure network-level detection methods, the physical layer of wireless communication can provide additional fingerprint features that are extremely difficult for attackers to forge. The emerging paradigm of cyber-physical trust leverages these physical-layer features to establish trust relationships in agent communication. Radio frequency fingerprinting utilizes subtle process differences in the transmitter hardware to generate a unique and unforgeable signal signature, which is used to verify the identity of the signal source and distinguish legitimate agents from counterfeiters[44]. Channel state information and received signal strength exploit the fact that transmitters located at different physical positions cannot produce exactly the same channel profile; by comparing the expected spatial signature with the actual measured value, injection attacks or relay attacks can be detected[124]. When the agent is equipped with an antenna array, angle-of-arrival and distance measurements allow the estimated direction of the received message to be cross-checked with the known position of the claimed sender, thereby obtaining a location-based trust metric[44]. These mechanisms operate independently of control logic and are inherently robust to algorithmic deception, complementing model-based SSE and anomaly detection methods. Although their integration with distributed consensus protocols is still in its infancy, this direction holds great promise for building a bridge of defense in depth between the information domain and the physical domain[44,45].

For agricultural robot swarms, perception-layer defense must be co-designed with the physical characteristics of field sensing[10]. GNSS/RTK spoofing can be detected by cross-validating with visual odometry, LiDAR scan matching, and crop-row detection. Multispectral deception can be detected by comparing current NDVI maps with historical field maps and with independent thermal or RGB imagery. Camera spoofing can be detected by consistency checks between stereo disparity, LiDAR depth, and IMU motion. LiDAR/IMU replay can be detected by temporal consistency and challenge–response mechanisms[83]. Physical-layer trust

based on RF fingerprints and channel state information can verify the identity of neighboring robots and ground stations[44,124]. However, these mechanisms must be lightweight enough for embedded platforms and robust to dust, vibration, and temperature variations. Experimental validation should include field trials in orchards, greenhouses, and open fields, where GNSS availability, crop canopy, and weather conditions vary significantly.

From a cross-layer perspective, perception-layer defense cannot be designed in isolation[62]. The number of redundant sensors and the frequency of SSE directly affect the transmission load and the computation burden of the decision layer. For example, multimodal sensing in agricultural robot swarms improves spoofing detection but increases the amount of data to be transmitted over intermittent rural networks[99]. A cross-layer design should therefore jointly optimize sensor selection, transmission scheduling, and estimation accuracy under a unified resource budget. Conversely, SSE, attack detection filters, and physical-layer trust are primarily single-layer defenses acting on the perception node; when they are integrated with transmission scheduling or decision-layer observers, they become cross-layer defenses. For example, a secure estimator that adapts its sampling rate based on channel quality couples perception and transmission, whereas an observer-based detector that shares information with neighbors couples perception and decision-making[120].

A systematic comparison of perception-layer defenses reveals clear trade-offs. Compressive-sensing and ℓ_0/ℓ_1 methods assume sparse sensor attacks and require a centralized or strongly connected measurement structure; they provide exact recovery guarantees under redundancy conditions but scale poorly with network size and assume bounded attack sparsity[111,118]. SMT and set-based methods provide guaranteed resilient estimates under fewer than s attacked sensors, but their computational complexity grows combinatorially and they are difficult to distribute. Distributed observer-based schemes assume that the leader dynamics are detectable from local measurements and that the communication graph remains sufficiently connected; they scale better and provide asymptotic or exponential convergence, but their guarantees degrade under intermittent DoS and are sensitive to delay. MSR/W-MSR resilient consensus assumes an F-local or F-total Byzantine model and a graph-theoretic robustness condition [e.g., $(F+1)$-robustness]; it provides convergence to the convex hull of normal agents' initial values, but requires redundant connectivity and does not directly address FDI on the communicated states[14,85]. Physical-layer trust mechanisms based on RF fingerprints or channel state information assume that hardware imperfections or channel profiles are unforgeable; they provide identity verification with low computation, but their reliability degrades under high mobility, and their integration with control-layer stability proofs remains largely unexplored[44,124]. For validation, most of these methods are demonstrated only through numerical simulation, and only a small subset reports hardware-in-the-loop or field experiments. This gap is particularly pronounced for agricultural robot swarms, where multimodal sensing, dust, vibration, and temperature variation are rarely captured in simulation.

Although significant progress has been made in perception-layer defense of MASs, for truly resilient autonomous clusters to be deployed effectively, several deeply intertwined challenges still need to be addressed. A fundamental bottleneck lies in the difficulty of scaling provably SSE: most existing algorithms assume that centralized measurements or dynamic structures are simple enough[111]. Extending correctness guarantees to large-scale, heterogeneous, and nonlinear MAS without incurring unbearable communication and computational overhead remains an open issue. To make matters worse, under strong process noise and unknown inputs, it is also necessary to simultaneously detect and identify sensor anomalies. Distinguishing physical faults from deliberate attacks and locating compromised sensors are tasks that are naturally prone to false alarms, and the erosion of system performance by false alarms is no less than that by missed alarms[46,53]. The threat landscape has become increasingly severe due to covert and coordinated adversaries, who inject

time-varying and cross-agent correlated attack sequences that can simultaneously contaminate multiple sensors right under the noses of traditional residual monitors. To resist such threats, it is necessary to rely on a game-theoretic framework that can explicitly depict the collaborative adaptation of the opposing sides[113,125]. Meanwhile, physical-layer trust mechanisms based on radio frequency fingerprints or channel state information have not yet been convincingly verified on highly mobile agents, because these mechanisms rely on static or slowly changing channel conditions, which often do not hold in fast-moving dynamic scenarios[44]. Looking ahead, a key frontier is cross-layer collaborative design with a theoretical basis and integration across perception-layer detection, transmission-layer scheduling, and robust control at the decision-making layer, moving away from isolated single-layer optimization toward a resilient, integrated, and interconnected architecture.

3.2. Transmission-layer defense: communication resilience under adversarial disruption

This subsection reviews transmission-layer defenses against DoS, FDI, replay, and asynchronous multi-channel attacks. We distinguish single-layer communication scheduling from cross-layer co-design of event-triggering, quantization, topology switching, and control.

The transmission layer is the nervous system of an MAS. It enables agents to exchange state information, synchronize their decisions, and coordinate their actions. As modeled in Section 2.3, transmission links are the primary targets of DoS attacks and FDI attacks on communicated data. When a communication link from agent j to agent i is under DoS attack, the received signal is erased [$\delta_{ij}(t) = 0$]; when under FDI attack, it is corrupted by an additive or multiplicative signal $\delta_{ij}{}^{a}(t)$. The defense at the transmission layer must simultaneously meet the interwoven requirements of availability and integrity: it is necessary to ensure that critical information can still reach its destination under intermittent channel congestion, while also preventing forged data from contaminating the coordination process. One often underestimated, but equally crucial requirement is efficiency - communication itself consumes energy and bandwidth, and excessive transmission directly expands the attack surface.

Event-triggered control is a rather mature technology in sampled-data systems. Its core idea is to transmit data only when the preset trigger conditions are violated[126-128]. In the context of DoS attacks on MASs, event-triggered control plays a dual role: first, it reduces the number of transmission attempts, thereby lowering the probability that transmissions encounter interference windows; second, it can flexibly adjust the transmission interval to the system's current needs. Even if packet loss occurs, it can still maintain considerable performance. For an agent i, the classical event-triggered condition is

$$t_{k+1}^{i} = \inf \left\{ t > t_{k}^{i} \mid \|e_i(t)\| \geq \sigma \|z_i(t)\| \right\}, \tag{31}$$

where $e_i(t) = x_i(t_k{}^i) - x_i(t)$ is the measurement error, $z_i(t)$ is a suitable local state-dependent signal (e.g., the consensus error), and $\sigma > 0$ is a design threshold. Under DoS attacks, the condition (31) must be augmented to account for the fact that a triggered transmission may be blocked. Two complementary strategies have been proposed.

If the communication channel is interfered with at the moment of trigger, the transmitted data will be temporarily cached. Once the attack stops, the system retries the transmission immediately. To ensure stability, it is usually mandatory to stipulate a retention time or a maximum allowable event interval[54]. Xu *et al.* designed an adaptive fuzzy event-triggering scheme whose triggering threshold can be dynamically adjusted according to the estimated duration of the attack, thereby ensuring that the controller can receive updated information within a bounded period of time after the DoS attack ends[31].

Dynamic event-triggered control (ETC) introduces an auxiliary internal variable $\eta_i(t)$ that evolves according to

$$\dot{\eta}_i(t) = -\beta\eta_i(t) + \sigma\|z_i(t)\|^2 - \|e_i(t)\|^2, \tag{32}$$

with $\beta > 0$, and the triggering rule becomes $t_{k+1}^i = \inf\{t > t_k^i \mid \eta_i(t) + \theta(\sigma ||z_i(t)||^2 - ||e_i(t)||^2) \leq 0\}$, where $\theta \in (0, 1)$. Dynamic event-triggered mechanism (DETM) can further reduce the number of events compared to static ETC and provides a unified Lyapunov analysis framework[129,130]. Several recent works have merged DETM with DoS resilience: Yang *et al.* propose a self-triggered variant where the next triggering instant is predicted using only locally available information and the current attack status[41], while Zhou *et al.* integrate DETM with pinning control for cluster consensus on signed graphs under DoS[131].

A critical requirement for any ETC scheme is avoiding Zeno behavior, i.e., an infinite number of events in finite time. Under DoS, this is particularly challenging because communication blackouts may force events to accumulate at the attack boundaries. Most works prove a strictly positive minimum inter-event time (MIET) by bounding the evolution of $||e_i(t)\}| / ||z_i(t)||$ during both attack and attack-free intervals[29,132].

The sequential scaling attack considered by He and Mo[58] provides an instructive case study for transmission-layer defense. Unlike DoS, which erases signals, the scaling attack preserves the communication link but multiplicatively corrupts the transmitted data. The event-triggered protocols in[58] reduce transmission opportunities and thereby limit the number of corrupted packets, while the consensus condition explicitly relates the triggering threshold to the attack duration and frequency. This is a cross-layer transmission–decision defense: the transmission node is protected by reducing exposure, and the decision node is protected by a consensus protocol that tolerates bounded scaling. However, the approach assumes linear agent dynamics and a known scaling structure, and it does not address replay, Byzantine, or actuator attacks. These limitations motivate the broader attack taxonomy and the balanced defense comparison developed in this survey.

Self-triggered control eliminates the need to continuously monitor trigger conditions. Its core idea is that whenever an event occurs, the next transmission moment[133,134] is pre-calculated based on the current state and the model of the controlled object. Under DoS attacks, one of the major advantages of self-triggered control lies in the fact that it completely eliminates the need for continuous perception[126,127,133,134]; continuous perception itself is a high-energy-consuming operation and will also expose more attack entry points for attackers[97-100,135]. A typical self-triggered law for agent i takes the form

$$t_{k+1}^i = t_k^i + \Gamma_i\left(x_i(t_k^i), \hat{x}_{\mathcal{N}_i}(t_k^i), \theta_{\mathcal{N}_i}(t_k^i)\right), \tag{33}$$

where $\Gamma_i(\cdot)$ is a bounded positive function that depends on the agent's state, the last received neighbor states, and the DoS status $\theta_{\mathcal{N}_i}$ on its incoming links. For satellite formations under asynchronous DoS attacks, researchers have designed a self-triggered formation control scheme whose core lies in carefully designing event intervals to ensure that information from neighbors remains fresh enough to support consensus even in the event of limited communication losses[136]. As an alternative technical approach to event-triggered and self-triggered schemes, periodic sampling with variable sampling rates has also attracted attention. Deng and Wen[29] adopted a deliberately aperiodic, time-varying sampling sequence, making it difficult for attackers to synchronize interference with the sampling pattern.

When communication resources are severely limited, quantized control further compresses the transmitted information. Instead of full-state vectors, agents exchange quantized versions $\hat{x}_i = q(x_i)$, where $q(\cdot)$ is a uniform or logarithmic quantizer with finite precision[137,138]. The interaction between quantization and DoS

attacks creates a compounded information loss: packets may be lost due to jamming, and those that arrive carry quantization errors. The overall system can be analyzed using input-to-state stability (ISS) concepts, where both quantization error and packet dropout are treated as bounded disturbances[24,139]. A uniform quantizer with an event-triggered consensus protocol is integrated, showing that consensus to a bounded ball can be achieved provided the product of quantization resolution and maximum DoS duration is below a computable threshold[121]. More recently, zoom-in/zoom-out quantization[137] and dynamic quantization[140] have been explored to adapt the quantization level based on the attack intensity, thereby gradually refining the control precision during long attack-free intervals.

Since DoS attacks remove edges from the communication graph, the effective topology becomes time-varying: $\mathcal{G}(t) = (\mathcal{V}, \mathcal{E}(t))$ where $\mathcal{E}(t) = \{(j,i) \mid \theta_{ij}(t) = 1\}$. Stability analysis under such switching topologies typically employs ADT or switching frequency conditions[141,142]. A common sufficient condition for consensus is that the union graph over uniformly bounded time intervals contains a spanning tree[67]. If the DoS attacks satisfy the frequency–duration constraints (15) and (16), one can derive explicit bounds on the allowable duty cycle η that guarantee this union graph condition[24,81].

In addition to passively tolerating link removal, some methods actively reconstruct the communication topology to bypass attacked links. For instance, one study proposes a scheme that switches between backup topologies, driven by a continuous packet-loss monitoring function[112]. As long as at least one backup topology remains connected within each attack window, consensus can be preserved without restricting the frequency or duration of DoS. Similarly, Dai *et al.* embedded topology handover into the impulsive control framework, synchronizing the pulse time with the handover event, thereby achieving leader-following synchronization under multi-channel independent DoS attacks[34].

When the attacker is resource-constrained (e.g., limited jamming energy), the interaction between the defender (MAS) and the attacker can be modeled as a game. The defender chooses a transmission schedule or power allocation, while the attacker decides which links to jam and when. Let $\mathcal{P}_D$ and $\mathcal{P}_A$ be the action sets of the defender and attacker, and let $J(u_D, u_A)$ be the cost function (e.g., tracking error plus transmission energy plus attack energy). A Nash equilibrium (u_D^*, u_A^*) satisfies

$$J(u_D^*, u_A^*) \le J(u_D, u_A^*), \quad \forall u_D \in \mathcal{P}_D, \qquad J(u_D^*, u_A^*) \le J(u_D^*, u_A), \quad \forall u_A \in \mathcal{P}_A. \tag{34}$$

Zero-sum game formulations, where $J_D = -J_A$, have been used to design optimal jamming and anti-jamming strategies in wireless MASs[125,143]. Extensions to Stackelberg games, where the defender acts as the leader, yield proactive defense policies that anticipate the attacker's best response[113]. At the physical layer, dynamic power control based on the signal-to-interference-plus-noise ratio (SINR) provides an additional knob for mitigating jamming. By increasing transmission power during jamming periods or frequency-hopping, agents can maintain a minimum data rate at the cost of higher energy consumption[124].

While DoS attacks have received extensive attention at the transmission layer, FDI, replay, and Byzantine attacks require equally systematic treatment. For FDI on communication links, defenses include authentication, message authentication codes, and encryption, but these alone cannot prevent compromised insiders from injecting validly signed false data[23]. Control-aware detection is therefore needed, such as watermarking, dynamic challenge–response, and cross-checking with physical invariants[53]. For replay attacks, time-stamping, nonce-based protocols[83], and physical-layer fingerprints are effective[44], but they must be lightweight enough for resource-constrained robots. For Byzantine attacks, the transmission layer must support robust aggregation and trust management, so that compromised nodes cannot dominate consensus[14,85]. Network coding and redundant multi-path transmission can further improve resilience by

allowing the receiver to reconstruct the true signal even when some paths are corrupted[34,112]. These mechanisms complement DoS-oriented scheduling and together provide a more balanced transmission-layer defense.

Extreme farmland channel loss imposes unique constraints on transmission-layer defense[10]. Rural wireless links are intermittent, long-range, and obstructed by terrain, crop canopy, and farm buildings. Dust, rain, and high temperature cause additional fading and hardware degradation. Under such conditions, event-triggered and self-triggered communication are necessities rather than mere efficiency tools, but their thresholds must adapt to time-varying channel quality and task criticality[31,99]. During harvesting, position and task-completion updates are safety-critical and should be transmitted with higher priority and redundancy; during monitoring, lower-rate updates may suffice. Quantized communication can reduce bandwidth, but the quantization resolution must be co-designed with the control law to avoid instability[121,137]. Topology switching and UAV relaying can restore connectivity when ground links fail, but they introduce new attack surfaces and require trust management[34,112]. Game-theoretic power control and anti-jamming strategies must account for the energy constraints of battery-powered field robots and the limited spectrum available in rural areas[113,125]. Field validation should measure packet loss, delay, and jamming resilience under realistic agricultural conditions.

Transmission-layer defense is intrinsically coupled with the decision layer. Event-triggered and self-triggered schemes reduce communication and exposure, but the resulting intermittent information may degrade the stability and performance of resilient consensus or learning-based control. Conversely, decision-layer adaptation can tolerate some packet loss, but only within bounds that depend on the transmission schedule. This coupling motivates cross-layer co-design of triggering thresholds, quantizer resolution, and controller gains, rather than separate tuning. The hierarchical federated learning framework in[62] exemplifies such co-design by jointly considering communication efficiency, privacy preservation, and detection accuracy. In terms of classification, event-triggered, self-triggered, and quantized communication are single-layer defenses when they only schedule or compress transmissions; they become cross-layer defenses when their triggering thresholds, quantization resolution, or topology switching are co-designed with the control law or the estimator. For instance, an event-triggered observer-based controller couples transmission and decision-making, whereas a topology-switching scheme coordinated with impulsive control couples transmission and execution.

Transmission-layer defenses can be compared along the same six dimensions. Static event-triggered control assumes a known Lyapunov function or a known bound on the consensus error; it addresses DoS by reducing transmission opportunities and guarantees a positive MIET, but its threshold must be tuned conservatively, and its stability proof typically assumes that the union graph over bounded intervals contains a spanning tree[126]. Dynamic event-triggered control introduces an auxiliary internal variable and provides a unified Lyapunov framework with fewer events, but requires additional state and computation on each agent[129,130]. Self-triggered control eliminates continuous monitoring and is therefore attractive for energy-constrained platforms; however, it requires an accurate local model to predict the next triggering instant, and its guarantees degrade under model mismatch[133,134]. Quantized communication assumes quantization error is a bounded disturbance and provides ISS-type guarantees, but the quantization resolution must be co-designed with the control gain; otherwise, the closed-loop system may exhibit limit cycles[137,139]. Topology-switching and backup-topology schemes assume at least one connected backup graph within each attack window; they provide consensus without restricting DoS frequency or duration, but require additional coordination and may introduce switching transients[34,112]. Game-theoretic power control assumes a rational, energy-constrained adversary and yields Nash or Stackelberg equilibria; it provides proactive defense but requires knowledge of the adversary's cost function, which is often unavailable in practice[113,125]. Regarding

validation, most works use numerical simulation with up to five agents[97,98]; large-scale wireless testbeds with realistic fading, interference, and hidden terminals are rare, and field validation in agricultural settings is almost absent.

Despite the extensive body of work surveyed above, transmission-layer defense in MASs continues to confront several deeply intertwined challenges that hinder the deployment of resilient autonomous networks. A primary limitation lies in the oversimplified modeling of DoS attacks: most existing guarantees rest on either global jamming that simultaneously blocks all links, or independent link failures governed by known stochastic statistics[24,49]. The practically critical case of asynchronous, coordinated attacks that dynamically corrupt a time-varying subset of links remains largely unresolved[27]. This gap is compounded by the persistent separation of communication and control design. Event triggers, quantizers, and power levels are typically tuned in isolation from the control law, introducing inherent conservatism; a truly integrated co-design framework that jointly optimizes control performance and communication resource expenditure under security constraints is still lacking[62]. This separation crystallizes into a fundamental security–performance–resource trilemma, where attack resilience, control fidelity, and the consumption of energy, bandwidth, and computation are intrinsically coupled. Formal quantification of this trade-off and derivation of Pareto-optimal operating points remain open problems. The threat surface is further amplified by intelligent, protocol-aware adversaries that exploit detailed knowledge of event-triggered or self-triggered schemes to launch stealthy DoS attacks precisely at transmission instants, maximizing disruption with minimal energy[32]. Countering such attackers requires unpredictable scheduling or randomized transmission policies, but robustness guarantees remain nascent. Finally, a critical gap remains between theoretical proposals and practical validation: most schemes are demonstrated only on small-scale numerical examples involving at most five agents. Their scalability to hundreds or thousands of agents, and their performance on real wireless testbeds afflicted by non-ideal channel effects such as fading, interference, and hidden terminals, have yet to be convincingly demonstrated.

3.3. Decision-layer defense: learning and adaptive control for online evolution in unknown environments

This subsection reviews learning- and adaptive-control-based decision-layer defenses. We compare adaptive control, fuzzy/neural control, adaptive dynamic programming (ADP)/reinforcement learning (RL), and MARL in terms of required model knowledge, stability certification, sample efficiency, and adversarial vulnerability.

When the attack pattern, system dynamics, or environment is uncertain or time-varying, pre-designed fixed-gain controllers may fail to maintain the required performance and safety margins. Adaptive control and learning-based control provide two complementary paradigms that enable the MAS to evolve its defense strategy online by extracting knowledge from real-time data. In the four-node information flow architecture [Figure 3], these techniques primarily operate at the decision node (D) and the execution node (E), augmenting the nominal control law with mechanisms that compensate for unknown attack-induced perturbations or actuator faults, and even learning optimal countermeasures through interaction with the adversary.

The situation that adaptive control needs to deal with is: when the parameters or attack signals of the controlled object are unknown constants or change slowly, how to maintain the stability and tracking performance of the system[69,144]. In MASs, a typical adaptive security control scheme adds an attack amplitude estimator and a corresponding offset term to the nominal consensus or formation protocol. Consider agent i with dynamics $\dot{x}_i = Ax_i + B(u_i + \delta_i^a)$, where $\delta_i^a(t)$ is an unknown attack signal injected at the actuator (execution node). If δ_i^a can be parameterized as $\delta_i^a = \Phi_i^\top(t)\theta_i^*$ with known basis function $\Phi_i(t)$ and unknown constant parameter θ_i^*, an indirect adaptive controller takes the form

$$u_i = -Kx_i - \Phi_i^\top(t)\hat{\theta}_i, \tag{35}$$

where $\hat{\theta}_i$ is updated online via a Lyapunov-based adaptation law, e.g.,

$$\dot{\hat{\theta}}_i = \Gamma\Phi_i(t)B^\top Pe_i, \tag{36}$$

with $\Gamma = \Gamma^\top > 0$ the adaptation gain, P the solution to a Lyapunov equation, and e_i the tracking error[77]. Stability is guaranteed through a quadratic Lyapunov function $V = e_i^\top Pe_i + \tilde{\theta}_i^\top\Gamma^{-1}\tilde{\theta}_i$, where $\tilde{\theta}_i = \hat{\theta}_i - \theta_i^*$. This framework has been extended to heterogeneous MASs under DoS and FDI attacks by Zhang *et al.* and Jiang *et al.*, where the adaptive laws also incorporate projection operators to keep estimates within physically admissible bounds[36,37].

For nonlinear MASs, the strict feedback form Equation (7) is naturally suitable for the application of adaptive backstepping methods. In each step of the backstepping design, the virtual control law is supplemented with an adaptive term to counteract the disturbances caused by unknown nonlinearity or attacks. Xu *et al.* combined adaptive fuzzy backstepping with an event-triggering mechanism to handle actuator failures and DoS attacks simultaneously[31]. In the containment control problem, adaptive NNs are used to approximate unknown nonlinear functions and compensate for the loss of neighbor information caused by DoS attacks[145]. One key challenge in adaptive security control is parameter drift: during a long DoS attack interval, the adaptive law may receive no new information, causing outdated errors to accumulate and potentially destabilizing the system. Common solutions include sigma correction, dead-zone correction, and parameter projection to prevent the estimated value from growing unbounded[144].

When the attack δ_i^a or the system nonlinearities $f_i(x_i)$ cannot be linearly parameterized, universal function approximators such as fuzzy logic systems (FLSs) and NNs provide a powerful alternative[146,147]. An FLS or NN can be written as

$$f(x) = W^{*\top}\Phi(x) + \epsilon(x), \tag{37}$$

where W^* is an unknown ideal weight matrix, $\Phi(x)$ is a vector of basis functions (fuzzy membership or activation functions), and $\epsilon(x)$ is a bounded approximation error. The actual control employs an estimate $\hat{W}$, updated online via gradient-based or Lyapunov-based rules.

In secure MASs, FLSs and NNs have become indispensable function-approximation tools, mainly supporting resilience against malicious attacks in three ways. First, they approximate the unknown attack signal δ_i^a or its envelope, thereby introducing compensation terms in the controller to actively counteract injected perturbations[36]. Second, when there is a lack of an exact dynamic model of the agent, these approximators can directly learn the ideal control law in a model-free adaptive control framework, thereby bypassing the explicit system identification step[148]. Third, they provide the infrastructure for learning value functions and optimal strategies for resilient control based on RL, acting as networks of critics and actors. A representative example of this third role is the event-triggered RL scheme designed for stochastic nonlinear systems[55], where critic and actor NNs are employed to approximate value functions and synthesize control policies, respectively.

While adaptive control compensates for known-structured uncertainties, RL and ADP aim to learn optimal defense policies from data without requiring detailed attack models[149,150]. The secure control problem is often framed as a two-player zero-sum game between the defender (controller) and the attacker (disturbance or adversary), leading to an H_∞-optimal control formulation:

Table 5. Comparison of learning and adaptive control paradigms for secure MASs

Paradigm	Core mechanism	Required knowledge	Security guarantee	Main limitations
Adaptive control	Lyapunov-based parameter update[77,144]	Structure of attack (basis functions)	Asymptotic stability, bounded signals	Parameter drift during DoS, requires linear parameterization
Fuzzy/NN control	Universal approximation + gradient update[147]	None (model-free)	UUB	Approximation error, high computational load
ADP/RL	Value/policy iteration, temporal difference[149,150]	System dynamics or state-input data	H_∞ gain, optimality	Exploration–exploitation trade-off, training time
MARL	Decentralized policy gradient, Q-learning[45,152]	None (fully data-driven)	No formal guarantees in most cases	Scalability, non-stationarity, communication overhead

MASs: Multi-agent systems; DoS: denial-of-service; NN: neural network; UUB: ultimately uniformly bounded; ADP: adaptive dynamic programming; RL: reinforcement learning; MARL: multi-agent reinforcement learning.

$$\min_{u}\max_{d}\int_{0}^{\infty}\left(x^{\top}Qx+u^{\top}Ru-\gamma^{2}d^{\top}d\right)dt, \tag{38}$$

where d represents the attack signal and $\gamma > 0$ is the prescribed disturbance attenuation level. The solution to this game is the saddle point (u^*, d^*) satisfying the Hamilton–Jacobi–Isaacs (HJI) equation. ADP circumvents the curse of dimensionality by using a critic–actor architecture: (i) the critic NN approximates the optimal value function $V^*(x)$; (ii) the actor NN approximates the optimal control policy $u^*(x)$ [and sometimes the worst-case attack $d^*(x)$]. The weights of both networks are updated iteratively using temporal difference errors or integral reinforcement learning (IRL), which eliminates the need for knowledge of the system drift dynamics[151].

For MASs under cyber attacks, several recent works have exploited ADP. Xu and Wu proposed a data-driven off-policy RL algorithm to learn the completely unknown system matrix of the leader, achieving optimal formation control in the presence of DoS[38]. Nian *et al.* utilized ADP to solve constrained social cost minimization for heterogeneous MASs under DoS attacks with coupled cost functions[42]. A particularly promising direction is MARL for cybersecurity[45]. In this framework, each agent is an independent learner that adapts its policy based on local observations, potentially including explicit models of adversarial agents. Cooperative MARL algorithms, such as multi-agent deep deterministic policy gradient (MADDPG) and QMIX, have been applied to detect and mitigate jamming, spoofing, and FDI attacks in a fully distributed manner, albeit often without formal convergence guarantees.

Although the preceding discussion has outlined the main learning and adaptive paradigms, a deeper critical analysis is necessary because these methods are increasingly presented as general solutions to secure control. We therefore examine four fundamental challenges: stability certification, sample efficiency, exploration safety, and vulnerability of learning components to adversarial manipulation, and compare learning-based methods with model-based resilient control on a common basis. Table 5 summarizes the principal paradigms and their practical considerations.

Model-based resilient control typically provides Lyapunov-based asymptotic stability, ISS, or finite-time convergence guarantees under explicit attack bounds[24,69,153]. Observer-based resilient consensus and SMC can certify stability under bounded DoS frequency and duration, and adaptive backstepping can guarantee ultimately uniformly bounded (UUB) signals under parameterized uncertainties[37,77]. In contrast, RL and ADP provide stability guarantees only under restrictive conditions: either the value function is known to be Lyapunov, or the policy is constrained by a control barrier function (CBF) or neural barrier certificate[110,149]. Most deep RL algorithms, including deep deterministic policy gradient (DDPG), proximal policy optimization (PPO), and soft actor-critic (SAC), offer no stability guarantee during training, and their

convergence is asymptotic in probability at best[45,152]. MARL is even weaker: non-stationarity induced by simultaneous policy updates invalidates the Markov assumption underlying single-agent convergence proofs, and global convergence is generally not guaranteed[152,154]. Learning-based secure control therefore cannot currently replace model-based resilient control in safety-critical applications; at best, it can complement it.

Model-based resilient control requires a system model but needs little or no online data. Adaptive control requires only parameter identification data and converges in the order of the number of unknown parameters. In contrast, deep RL and ADP typically require thousands to millions of interactions to learn a useful policy, and the sample complexity grows with state dimension, action dimension, and network size[159,151]. In multi-agent settings, the joint state-action space grows exponentially with the number of agents, making tabular or naive function-approximation approaches infeasible[152]. Although transfer learning, meta-learning, and digital-twin-based pre-training can reduce online samples, they assume access to a high-fidelity simulator or a related task, which is often unavailable for adversarial scenarios[155,156]. Sample efficiency therefore remains a fundamental bottleneck for deploying RL-based secure control on physical platforms.

Model-based resilient control separates design and operation: the controller is certified offline and then executed. RL, by contrast, learns through trial and error, and the exploration phase is inherently unsafe. An agent may take actions that violate physical constraints, damage actuators, or cause collisions. Safe RL with CBFs or shielding can mitigate this, but the shield itself must be certified, and its interaction with the learning policy is not fully understood[110]. In MASs, exploration safety is further complicated because one agent's unsafe action can propagate to neighbors. This is particularly problematic for agricultural robot swarms, where an exploratory action of a sprayer or weeder can directly damage crops or contaminate soil[10]. Existing safe RL methods have not been convincingly validated on such safety-critical heterogeneous platforms.

Learning-based controllers introduce new attack surfaces that model-based designs lack. The critic network can be poisoned through manipulated rewards or falsified value targets; the actor network can be misled by adversarial perturbations in the observation space; and the replay buffer can be corrupted by injecting fabricated transitions[154]. A small, carefully crafted perturbation that is imperceptible to residual-based detectors can cause the policy to output unsafe actions. Federated learning, often used to preserve privacy, is vulnerable to Byzantine participants that upload poisoned model updates[157,158]. In contrast, model-based resilient control has a well-defined attack surface (the measurements and the communication links) and its stability proofs often explicitly account for bounded adversarial inputs. This asymmetry means that learning-based defenses may be circumvented by adversaries who target the learning process itself rather than the control loop.

On a common basis, model-based resilient control excels in stability certification, interpretability, and verifiability, and it requires little online data; its main limitations are conservatism, dependence on a reasonably accurate model, and difficulty in adapting to unmodeled or time-varying attacks. Learning-based control excels in adaptability, model-free operation, and handling of high-dimensional nonlinear dynamics; its main limitations are lack of formal guarantees, high sample complexity, unsafe exploration, and new adversarial vulnerabilities. These are complementary rather than substitutable strengths. A promising direction is hybrid architectures in which a model-based resilient controller provides a certified safety baseline, while a learning component improves performance and adapts to unknown attacks within the certified envelope. Examples include safe RL with CBFs, ADP with Lyapunov-based constraints, and learning-augmented observers[62,110,151]. However, the theoretical foundations of such hybrids - especially formal stability and security guarantees under adversarial injection - remain largely open.

Another gap concerns deploying learning-based and adaptive methods on real physical platforms. Computational and memory constraints: most deep reinforcement learning (DRL) and MARL algorithms require high-dimensional NNs, large replay buffers, and frequent gradient updates, which exceed the onboard computation and memory of embedded platforms such as UAVs, agricultural ground robots, and low-power IoT devices[10,99]; edge offloading and network pruning are promising remedies but introduce additional communication delays and new attack surfaces. Real-time and communication constraints: learning-based controllers often rely on frequent information exchange for policy evaluation, gradient aggregation, or federated model updates, and the resulting delays and packet losses in intermittent wireless networks can destabilize the learning process and invalidate stability proofs that assume perfect or bounded-delay communication. Robustness of security boundaries: NN policies are vulnerable to adversarial examples, data poisoning, and reward manipulation, and the exploration phase is inherently unsafe; although safe RL with CBFs and neural barrier certificates has been proposed, formal guarantees under active adversarial attacks are still largely missing[110]. Simulation-to-reality gap: real platforms exhibit unmodeled dynamics, actuator nonlinearities, sensor noise, calibration errors, communication delays, and hardware aging, which can significantly degrade a policy that performs well in simulation; hardware-in-the-loop testing, digital twins, and domain randomization are necessary but rarely reported in the secure control literature[155,156]. Agricultural robot swarm case: onboard computation is limited by cost, weight, and power; rural communication is intermittent and lacks redundancy; multimodal sensors are exposed to dust, rain, vibration, and extreme temperatures; and actuator attacks on sprayers, weeders, and wheel motors can directly affect crop yield and food safety[10]. Cross-layer co-design, lightweight and verifiable learning architectures, and extensive field trials are therefore indispensable before deployment.

Decision-layer defenses such as resilient consensus, observer-based control, adaptive control, and RL are single-layer defenses when they only process the information available at the decision node. They become cross-layer defenses when they actively shape perception (e.g., by requesting additional measurements), transmission (e.g., by adapting communication schedules), or execution (e.g., by respecting actuator constraints). Safe RL with CBFs, for example, couples decision-making and execution[110]; federated anomaly detection couples decision-making and transmission[157,158]. These cross-layer architectures are essential for handling hybrid and stealthy attacks. For agricultural embodied intelligent robot swarms, decision-layer defense must additionally integrate task-level semantics with security: learning-based policies can adapt to unknown soil conditions, crop variability, and weather changes, but they must respect hard physical constraints such as sprayer flow limits, weeder torque bounds, and safe distances to humans and animals[10]. Experimental studies should evaluate task-level performance, such as spraying accuracy, weeding completeness, and harvest efficiency, under adversarial conditions.

Learning-based and adaptive defenses should also be balanced across attack classes. For FDI attacks, adversarial training and robust regression can improve the detector's resilience to stealthy injections, while watermarking can be integrated into the learning objective. For replay attacks, temporal consistency learning and sequence-based anomaly detection can distinguish replayed signals from genuine ones[83]. For Byzantine attacks, robust federated learning and Byzantine-resilient aggregation rules can prevent malicious agents from poisoning the global model[14,157]. For actuator attacks, safe RL with CBFs can constrain the policy to physically safe actions even under adversarial actuation[110,115]. These directions are distinct from DoS-oriented event-triggered learning and require different assumptions, loss functions, and safety certificates. A balanced research agenda should therefore develop learning-based defenses for the full spectrum of attack classes, not only for DoS.

3.4. Execution-layer defense: fault-tolerant actuation under physical attacks

This subsection reviews execution-layer defenses against actuator attacks and faults. We compare passive fault-tolerant control (FTC), active FTC, SMC, Nussbaum-function compensation, anti-windup design, and impulsive control.

The execution layer is the final link in the four-node information flow architecture [Figure 3], and it is here that the control instruction u_i calculated by the decision-making layer is physically imposed on the actuator. Even if the security at the perception, transmission, and decision-making levels is all foolproof, a successful attack or a hardware failure at the execution level is sufficient to directly change the force, torque, or voltage that drives the robot, rendering all the resilience measures upstream futile. Actuator anomalies can be described by the fault model (11) and the saturation model (12).

3.4.1. Fault-tolerant control

Fault-tolerant control aims to preserve acceptable system performance in the presence of actuator faults or attacks, typically by exploiting hardware or algorithmic redundancy and online reconfiguration. Two complementary design philosophies have been established. Passive FTC treats actuator faults as a class of admissible uncertainties and relies on a fixed robust controller designed to tolerate a predefined fault envelope. A representative formulation employs H_∞ synthesis, where the plant is augmented with a worst-case fault model and the controller minimizes the energy gain from fault and disturbance inputs to the regulated output[159]. In the MAS setting, multiplicative faults have been cast as time-varying parametric uncertainties, and a distributed controller with sufficient gain margins can maintain consensus without explicit fault detection[153], albeit at the cost of inherent conservatism, since the design must cover the full anticipated fault range and may sacrifice nominal performance. Active FTC, by contrast, incorporates a fault detection and diagnosis (FDD) module that explicitly identifies the fault type and magnitude before reconfiguring the controller[114]. For agent i, the FDD module produces online estimates $\hat{\rho}_i(t)$ and $\hat{\phi}_i(t)$ of the multiplicative and additive fault parameters, typically through observer-based residual generation or data-driven classifiers, and the control law is reconfigured as

$$u_i(t) = \frac{1}{\hat{\rho}_i(t)}\left(u_i^{\mathrm{nom}}(t) - \hat{\phi}_i(t)\right),$$

where u_i^{nom} denotes the nominal control signal. In distributed MASs, reconfiguration must be coordinated across agents: the loss of actuation capability in one agent may require its neighbors to adjust their control efforts to compensate for the resulting capability gap. This principle has been demonstrated by an integrated scheme combining a distributed fault estimator and a switching controller, which is designed to handle simultaneous DoS attacks and actuator faults in heterogeneous MASs[160].

3.4.2. SMC and anti-windup design

SMC is a natural candidate for execution-layer defense due to its inherent insensitivity to matched uncertainties and its finite-time convergence properties[161,162]. In the context of MASs under actuator attacks, the attack signal $\delta_i^a(t)$ in Equation (18) is often matched through the input channel, i.e., $\dot{x}_i = f_i(x_i) + B_i(u_i + \delta_i^a)$. A classical sliding mode controller forces the system state onto a sliding surface $\mathcal{S} = \{x \mid \sigma(x) = 0\}$ in finite time, and the equivalent control on the surface cancels the attack signal provided its magnitude is bounded by a known function.

Specifically, consider the sliding variable $\sigma_i(x_i)$ and the reaching law $\sigma_i = -K\mathrm{sgn}(\sigma_i)$. The control law

$$u_i = -(C_i B_i)^{-1}\left(C_i f_i(x_i) + K\,\mathrm{sgn}(\sigma_i)\right) \tag{39}$$

with C_i such that C_iB_i is nonsingular, ensures that once on the surface, the dynamics are independent of δ_i^a as long as $||\delta_i^a|| \le K$. Cheng *et al.* extended this to stochastic MASs under Markov-chain-driven DoS attacks, proving that the sliding mode is reachable in probability despite random packet losses[115]. Furthermore, adaptive SMC schemes can online-estimate the bound K, thus removing the need for a priori knowledge of the maximum attack magnitude[163].

Physical actuators invariably possess limited authority, modeled by the saturation function (12), and under malicious attacks this authority can be further constricted when an adversary injects a persistent bias that pushes the input to its limits, eroding the effective actuation margin. To preserve resilience under such constraints, two principal strategies have emerged. Anti-windup augmentation supplements the nominal controller with a compensator that activates upon saturation[79], but the scheme must be hardened so that saturation triggered by an attack, rather than by a legitimate large control demand, is correctly accommodated; Chen *et al.* achieved this by incorporating a low-gain feedback design that prevents saturation during DoS-induced transients in bipartite consensus[164]. When the actuator effectiveness is unknown and may vary erratically due to multiplicative deception attacks, Nussbaum-type functions provide a mechanism to handle the resulting unknown control direction. The control law assumes the form $u_i = N(\xi_i)\bar{u}_i$, where $N(\xi_i)$ is a Nussbaum function [e.g., $\xi^2\cos(\xi)$] and $\bar{u}_i$ is a nominal stabilizing signal; the function adapts its sign to compensate for unknown sign changes in the actuator gain[116]. This approach has been successfully deployed by Jiang *et al.* in adaptive secure control of nonlinear MASs, where Nussbaum functions accommodate unknown time-varying actuator faults induced by deception attacks, thereby maintaining stability despite severely degraded actuation authority[37].

3.4.3. Impulsive control

Impulsive control applies control actions only at a sequence of discrete time instants $\{t_k\}$, thereby minimizing the reliance on continuous communication and actuation[165,166]. In the four-node architecture, impulsive control acts at both the transmission and execution layers: since the control signal is piecewise constant and changes only at impulses, the system is less sensitive to short-term DoS attacks on the communication network, and the actuator is only active at sparse instants, reducing the window of vulnerability to execution-layer attacks.

The impulsive control law for agent i is typically of the form

$$\Delta x_i(t_k) = -c \sum_{j \in \mathcal{N}_i(t_k)} a_{ij}\left(x_i(t_k^-) - x_j(t_k^-)\right), \tag{40}$$

where c is the impulsive gain, $x_i(t_k^-)$ denotes the left limit, and $\mathcal{N}_i(t_k)$ is the set of neighbors whose information is available at the impulse instant. Dai *et al.* proposed a topology-switching-based impulsive control scheme for leader-following synchronization under hybrid deception and multi-channel independent DoS attacks[34]. By synchronizing impulse instants with backup-topology switching, the algorithm ensures that impulsive correction uses only uncorrupted information, even when the primary communication graph is partially blocked. Li *et al.* investigated resilient finite-time consensus for variable-order fractional-order MASs under node-based DoS attacks, where the impulsive control gain is adaptively tuned to achieve convergence within a prescribed time despite irregular attack patterns[40]. A key advantage of impulsive control is its inherent Zeno-freeness: by design, the impulse instants are chosen with a strictly positive lower bound, so the control signal never becomes dense in time.

Execution-layer defenses can be compared along the six dimensions introduced at the beginning of Section 3. Passive fault-tolerant control assumes a known fault envelope and provides robust stability without fault detection; it is simple and fast but conservative, and it may sacrifice nominal performance[76,114]. Active

fault-tolerant control assumes that faults can be detected and identified in time; it provides better nominal performance but relies on the accuracy of the fault diagnosis module, which may be degraded under simultaneous sensor attacks[114,160]. SMC assumes that the attack or fault is matched through the input channel and bounded by a known function; it provides finite-time convergence and insensitivity to matched uncertainties, but may cause chattering and requires a priori knowledge of the bound[161,162]. Adaptive SMC estimates the bound online, removing this requirement, but introduces additional adaptation dynamics[115,163]. Nussbaum-function-based control assumes unknown but constant or slowly varying control direction; it handles sign changes in the actuator gain but may produce large transient control effort[37,116]. Anti-windup design assumes that saturation can be detected and compensated; it preserves stability under input constraints but must distinguish attack-induced saturation from legitimate large commands[78,79]. Impulsive control assumes that corrections at discrete instants suffice for synchronization; it provides Zeno-free operation and reduces actuator exposure, but the impulse instants must be coordinated with communication availability, and the stability proof requires the impulsive interval to be bounded[34,165,166]. In terms of validation, most of these methods are evaluated on numerical examples or small-scale robotic testbeds; field validation on agricultural machinery under realistic actuator faults and adversarial attacks remains scarce[10].

Execution-layer defense in agricultural robot swarms must protect actuators that directly affect crop yield and food safety[10]. A compromised sprayer pump can cause over- or under-application of chemicals, leading to crop damage or chemical residues. A jammed weeder can leave weeds untreated or damage crops. A degraded wheel motor can cause uneven field coverage or collisions. These actuator anomalies may result from malicious attacks or genuine hardware faults, and distinguishing between them is difficult[46,53]. Fault-tolerant control strategies should redistribute tasks among neighboring robots, reconfigure motion plans, and adjust application rates[160]. Impulsive control can reduce actuator exposure by applying corrections only at discrete instants, which is particularly suitable for energy-constrained field robots[166]. The impulsive instants must be coordinated with communication availability and task deadlines. Cross-layer co-design of fault diagnosis, task allocation, and communication scheduling is essential for maintaining mission-level safety. Field experiments should include actuator fault injection, adversarial command injection, and hardware-in-the-loop testing on real agricultural machinery.

From a cross-layer perspective, execution-layer defense is coupled with perception and decision-making in a closed loop[46,62]. Actuator attacks or faults alter the physical state, which affects sensor measurements and may be misinterpreted as perception-layer anomalies. At the same time, fault-tolerant or impulsive control actions change the information available to the decision layer and may require additional communication. A cross-layer resilient architecture should therefore coordinate fault diagnosis, state estimation, communication scheduling, and control reconfiguration, rather than treating them as independent modules. In terms of classification, execution-layer defenses such as fault-tolerant control, SMC, Nussbaum-function compensation, anti-windup design, and impulsive control are single-layer defenses when they only compensate for actuator anomalies; they become cross-layer defenses when they are integrated with fault diagnosis (perception), communication scheduling (transmission), or task reallocation (decision-making). For example, a fault-tolerant controller that redistributes tasks among neighboring robots couples execution and decision-making[160], whereas an impulsive controller synchronized with communication availability couples execution and transmission[34]. These cross-layer mechanisms are particularly important for agricultural robot swarms, where actuator attacks can directly affect crop yield and food safety[10].

Although execution-level defense has advanced significantly, several deep-seated obstacles remain. Particularly challenging is how to distinguish malicious injection from real hardware failures - he manifestations of the two on the actuator are almost indistinguishable[53,76]. Real-time discrimination and the choice of reconfiguration, isolation, or counterattack based on this remains an extremely difficult detection

problem to this day. When both sensor and actuator channels are attacked simultaneously, this predicament becomes increasingly acute: this realistic scenario shatters the sensor cleanliness assumption that most active fault-tolerant control schemes rely on, and precisely reveals the lack of a principled cross-layer resilience framework[56]. Covert slow-drift attacks can quietly erode actuator health and often easily bypass traditional threshold-based fault-diagnosis modules. By the time they are detected, the system has often deviated from the safe operation envelope. At the engineering implementation end, advanced fault-tolerant and sliding mode algorithms typically involve matrix inversion and high-gain observers. These operations can easily exceed the computing endowment of low-cost embedded platforms. Therefore, algorithm lightweighting and hardware-in-the-loop verification have become indispensable steps. Ultimately, the true potential of the resilience of the execution layer can only be unleashed through joint design with the decision-making layer, such as enabling RL strategies to reach consensus while learning to avoid executor saturation[110,152]. However, this collaborative design paradigm is still in its infancy.

3.5. Cross-layer privacy-preserving secure control

A growing body of research combines privacy protection with security control in MASs. Differential privacy has been employed to mask the state of individual agents during consensus iterations to prevent adversaries from inferring private information from the exchanged messages[167]. Homomorphic encryption enables controllers to perform calculations directly on encrypted data, thereby achieving resilient coordination without revealing the original state[168]. Secure multi-party computation and federated learning are also applied to distributed anomaly detection, enabling agents to collaboratively identify attacks without sharing sensitive local data[169]. These methods introduce a fundamental trade-off between privacy levels and control performance: the noise injected for privacy protection can reduce estimation accuracy and may be exploited by adversaries.

Privacy-preserving mechanisms span all four information-flow nodes, and their coupling with resilience deserves explicit treatment[62]. Differential privacy is typically applied at the perception and transmission nodes to mask local measurements and transmitted states[167,170]; homomorphic encryption protects the decision node by allowing computation on encrypted data[168]; secure multi-party computation and federated learning couple transmission and decision-making for distributed anomaly detection[157,169]; and privacy-aware actuator commands protect the execution node. These cross-layer couplings create a fundamental trade-off between privacy level and resilience: noise injected for privacy can degrade estimation accuracy, slow consensus, and reduce detection sensitivity, while encryption and secure computation introduce latency and computational overhead[158,171]. Quantitatively, the privacy budget ϵ controls the amount of noise added: smaller ϵ yields stronger privacy but larger estimation error and slower convergence[167]. In MASs, this trade-off is further complicated because each agent's privacy mechanism affects the information available to its neighbors, and thus the overall resilience of the network[94,170]. The key open challenges are threefold. First, privacy mechanisms must remain robust when some participants are malicious and actively attempt to infer private information from the collaborative defense process[157,158]. Second, privacy budgets should adapt to real-time threat levels, relaxing when attacks are detected and tightening during normal operation. Third, formal stability and security guarantees must be established under privacy-induced disturbances, which are currently treated only as bounded noise in most analyses. This trade-off is particularly acute in agricultural robot swarms, where intermittent connectivity and limited onboard computation constrain the feasible privacy budget, and where regulatory requirements on data sharing may vary across jurisdictions[10,99].

4. FUTURE RESEARCH OUTLOOK

The comparative analysis in Section 3 reveals a consistent pattern across the four information-flow nodes. Methods with strong formal guarantees, such as observer-based secure estimation, resilient consensus under F-local models, and SMC, rely on restrictive assumptions of known attack bounds, sparse attacks, or

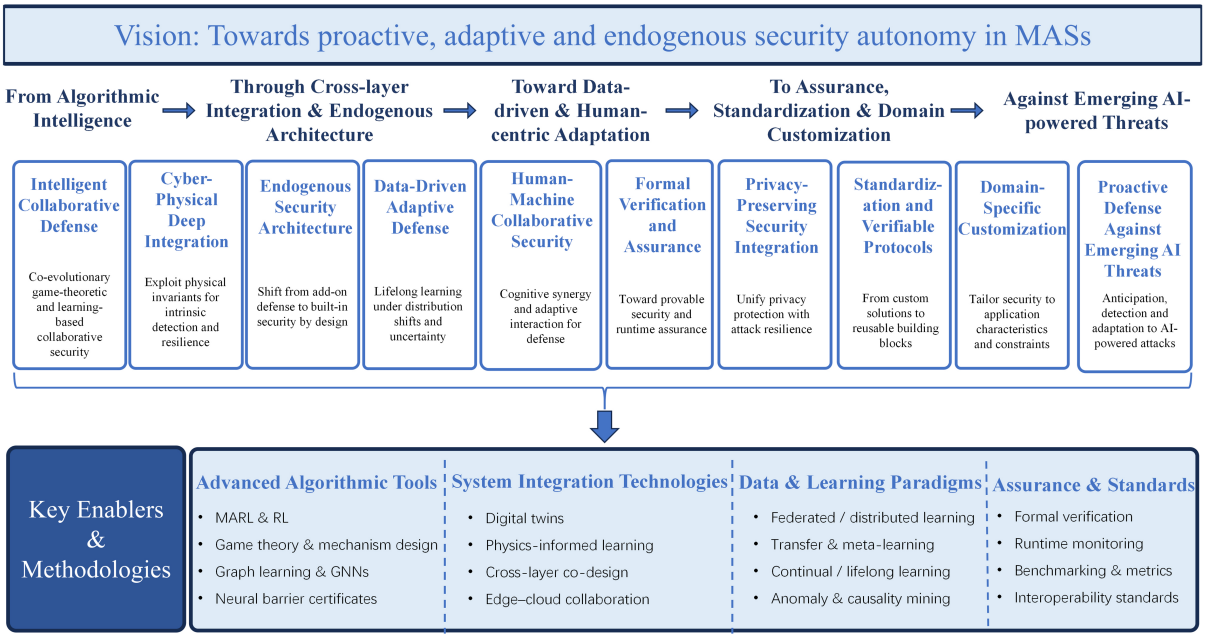


Figure 4. Framework of future research directions for secure MASs. MASs: Multi-agent systems; AI: artificial intelligence; MARL: multi-agent reinforcement learning; RL: reinforcement learning; GNNs: graph neural networks.

matched uncertainties. Methods with weaker assumptions, such as RL and MARL, offer greater adaptability but lack formal stability or security guarantees and remain difficult to certify. Communication requirements vary widely: event-triggered and quantized schemes reduce bandwidth but may degrade information freshness, whereas federated and distributed learning preserve privacy at the cost of frequent model updates. Computation ranges from lightweight scalar updates in adaptive control to high-dimensional NN training in deep RL. Validation remains dominated by numerical simulation, with hardware-in-the-loop testing and field trials rare, especially in agricultural settings. These observations define the concrete gaps that motivate the ten future research directions illustrated in Figure 4, each of which extends specific methods reviewed in Section 3. To ensure consistency, the ten directions in Figure 4 correspond exactly to the subsections below and the roadmap in Table 6. Table 6 at the end of this section provides a concise roadmap linking each direction to its corresponding gap and pathway.

4.1. Intelligent collaborative defense: from static protocols to co-evolutionary security

The resilient consensus and event-triggered schemes reviewed in Sections 3.2 and 3.3 assume fixed attack models and static trust, and the learning-based alternatives that could relax these assumptions still lack formal guarantees. This gap motivates a shift toward opponent-aware MARL, in which agents learn attack patterns and neighbor credibility through repeated interactions, and toward game-theoretic trust mechanisms that render cooperation incentive-compatible[152,154]. Privacy-preserving distributed learning can support such policies without sharing raw trajectories[170,172]. Three questions remain open: how to achieve provably stable trust evaluation under partial observability when malicious agents deliberately contaminate the learning process; how to guarantee safety during exploration (a challenge shared with data-driven adaptive defense); and how to scale collaborative defense to large-scale heterogeneous networks under communication and computation constraints. Neural barrier certificates offer a concrete pathway for the second question.

Table 6. Future research roadmap for secure MASs

Direction	Open challenge	Promising pathways	Horizon
Intelligent collaborative defense	Static trust, fixed attack models	Opponent-aware MARL, game-theoretic trust, safe RL[152,154,170,172]	Medium
Cyber-physical deep integration	Separation of cyber and physical detection	Physics-informed learning, invariant extraction[95,96,155]	Medium–long
Endogenous security architecture	Add-on rather than built-in security	Heterogeneous redundancy, moving target defense[63,108,175,176]	Long
Data-driven adaptive defense	No formal guarantees, catastrophic forgetting	Hybrid control, digital twins, lifelong learning[155,156]	Medium
Human-machine collaborative security	No situational reasoning in autonomous defense	Adjustable autonomy, explainable AI[109]	Medium–long
Formal verification and assurance	Simulation-only validation, state explosion	Compositional and runtime verification[110,179-182]	Long
Security-privacy co-design	Separate treatment of security and privacy	Federated detection, homomorphic encryption, differential privacy[157,158,170]	Medium
Standardization and verifiable protocols	Fragmented benchmarks and metrics	Standard taxonomies, open testbeds[183-185]	Short–medium
Domain-specific security-by-design	Generic solutions insufficient for diverse domains	Agricultural robot swarm security, UAV, smart grid, platoon[103,106,135,174]	Short–medium
Proactive defense against AI-powered threats	Fixed attack models, vulnerable learning components	Active deception, adversarially robust learning[188]	Medium–long

The time horizons are estimates based on current research trends and may shift as the field evolves. MASs: Multi-agent systems; MARL: multi-agent reinforcement learning; RL: reinforcement learning; AI: artificial intelligence; UAV: unmanned aerial vehicle.

4.2. Cyber-physical deep integration: exploiting physical invariants as intrinsic shields

Most detection and control methods compared in Sections 3.1 and 3.3 treat network attacks and physical dynamics as separate problems, and physics-based detection remains largely limited to linear time-invariant systems. Yet in cyber-physical MASs, attacks ultimately manifest as violations of physical invariants, conservation laws, kinematic constraints, stability margins, and consensus properties, which adversaries cannot tamper with and which therefore constitute reliable detection signals. Physics-informed learning can embed these invariants into the detection and control models, enabling identification of covert attacks that pure data-driven detectors miss[44,95,96,155,173]. Domain-specific constraints such as maximum acceleration, safe distance limits, and power flow equations extend this idea naturally to vehicle formations, UAV swarms, power systems, and agricultural robot swarms, where kinematic and safety constraints are tightly coupled with crop and terrain conditions[103,106,107,174]. Open questions include how to extract physical invariants from high-dimensional nonlinear dynamics without a complete first-principles model, how to design controllers that degrade gracefully when detection fails, and how to distinguish legitimate physical disturbances from attack-induced anomalies in noisy and uncertain environments.

4.3. Endogenous security architecture: designing security into system DNA

The defense taxonomy of Section 3 reveals that almost all reviewed mechanisms are add-on layers, and that even zero-trust approaches focus on access control rather than architecture refactoring. This patching paradigm carries three structural flaws: it imposes significant performance and communication overhead, it creates a single point of failure within the security layer itself, and it cannot resist zero-day vulnerabilities that exploit design flaws in the underlying control architecture. A shift toward native security design is therefore necessary, embedding authentication, redundancy, diversity, and moving target defense into the control architecture from the earliest design stage. Promising directions include dynamic heterogeneous redundancy, in which functionally equivalent but differently implemented controllers run in parallel and arbitrate their outputs; graph-theoretic topology design that maximizes inherent Byzantine resilience without additional security modules[108,175]; and moving target defense that dynamically varies communication modes,

control parameters, and topology to increase attacker uncertainty[63,176]. These extend the resilient consensus and topology-switching methods reviewed in Sections 3.2 and 3.3 from runtime reconfiguration to design-time guarantees. The open questions are how to quantify the intrinsic security level of a control architecture, how to balance redundancy and diversity against the resource constraints of UAV swarms and agricultural robots, and how to design composable mechanisms that support the construction of large-scale systems from proven components.

4.4. Data-driven adaptive defense: continual learning for evolving threats

The adaptive, fuzzy/neural, and ADP/RL methods compared in Section 3.3 share a common weakness: they are vulnerable to data poisoning, suffer catastrophic forgetting, and lack certifiable stability and security bounds, whereas model-based methods cannot adapt to unmodeled attacks[97,100,102,104,177,178]. Hybrid architectures that use robust control as a certified safety baseline and data-driven learning for performance improvement offer a natural way forward. Safe RL with neural barrier certificates can provide exploration-time guarantees, lifelong learning can prevent forgetting, and digital twins can provide a safe pre-training sandbox that addresses the scarcity of rare attack scenarios[155,156]. The open questions are how to certify a learning component that operates within a model-based safety envelope, how to detect poisoned or adversarial training data in real time, how to guarantee exploration safety in heterogeneous MASs where one agent's action affects others, and how to provide composable stability and security guarantees when perception, transmission, decision-making, and execution are each partially learning-enabled.

4.5. Human-machine collaborative security: adjustable autonomy for high-stakes decisions

Fully autonomous defenses, as reviewed in Section 3.3, respond rapidly but lack situational reasoning when confronted with novel attacks or situations outside the training distribution, and may make highly confident but disastrous decisions. Purely manual systems, by contrast, suffer from slow response, alarm fatigue, and cognitive bias. The complementary strengths of humans and automated systems, rapid consistent processing of known patterns *vs.* situational reasoning, novelty detection, and ethical judgment, are not yet exploited by existing human-in-the-loop methods, which focus on performance rather than security decisions[109]. Adjustable autonomy offers a principled framework, dynamically allocating authority between autonomous defenses and human operators based on threat severity, system confidence, and decision criticality. Explainable AI is essential for calibrating trust, particularly for large language model integrated agents whose reasoning may be opaque. The open questions are how to design interfaces that present complex attack situations without causing information overload, how to formally model human trust and automation bias, and how to ensure security during transitions between autonomous and manual control under time pressure.

4.6. Formal verification and assurance: building certifiably secure MASs

As noted in the synthesis above, the validation-method comparison in Section 3 shows that most reviewed algorithms are verified only through simulation and limited experiments, which cannot exhaust all possible attack scenarios or provide formal security guarantees. This constitutes a fundamental obstacle to deployment in safety-critical regulated domains such as automotive, aerospace, and industrial control. Although formal methods have been applied to specific control problems[110,179], the state explosion problem limits their application to large-scale distributed MASs, and most existing methods are designed for static deterministic systems rather than the attack randomness, event-triggered communication, and adaptive learning components that characterize modern secure control. Compositional verification, which decomposes global security properties into locally verifiable conditions on individual agents, extends the stability proofs of observer-based and resilient consensus methods reviewed in Sections 3.1 and 3.3 to larger systems[180]. Runtime verification supplements offline proofs by monitoring behavior after deployment, and probabilistic model checking handles random attack models and randomized defenses[181,182]. The open questions are how to scale compositional methods to large-scale heterogeneous networks, how to verify

learning-enabled defense components that existing frameworks treat as black boxes, and how to design specification languages that capture the dynamic nature of control requirements rather than borrowing general software security logic.

4.7. Security-privacy co-design: from trade-off to synergy

Security control and privacy protection are largely studied as independent problems, as the cross-layer discussion in Section 3.5 makes clear: security research promotes information sharing for attack detection, whereas privacy research restricts it to protect sensitive local data. This separation creates a fundamental tension, because effective collaborative defense requires sharing detailed status and anomaly information that may leak operational data, create new attack surfaces, or violate regulatory requirements. Federated anomaly detection, homomorphic encryption, and dynamic differential privacy can extend the privacy-preserving secure control methods reviewed in Section 3.5 to collaborative attack detection without disclosing raw data[94,157,158,170,171]. Semantic communication, which transmits task-related semantics rather than raw data, offers a particularly promising route to harmonizing security and privacy. The open questions are how to characterize the fundamental trade-off between privacy budget and resilience, how to design privacy mechanisms robust to adversarial participants who exploit the collaborative defense process, and how to integrate privacy with event-triggered and resource-constrained control without incurring the overhead of existing cryptographic methods.

4.8. Standardization and verifiable security protocols: from custom solutions to reusable building blocks

The comparative tables in Section 3 reveal that attack models and performance metrics vary widely across papers, preventing fair comparison and reproducibility, and that even widely studied problems such as DoS-resilient consensus lack common benchmarks. This fragmentation hinders industrial adoption, because practitioners cannot reuse well-tested security building blocks but must redesign solutions for each application. The field must therefore develop standardized security protocols, benchmark problems, and evaluation metrics analogous to transport layer security (TLS) in secure network communication. Key work includes general attack classification methods that support fair comparison, verifiable security protocols for consensus, formation control, and distributed optimization, and open-source reference implementations and testbeds[155,156]. Standardized security indicators that quantify resilience, overhead, and detection performance from multiple dimensions will provide the foundation for principled comparison[183-185]. The open questions are how to design protocols flexible enough for heterogeneous agents with different computation, communication, and energy constraints; how to develop benchmark problems that represent real attacks while remaining simple enough for controlled comparison; and how to enable third-party certification without disclosing proprietary control logic.

4.9. Domain-specific security-by-design: from generic solutions to application-aware resilience

The domain comparison in Section 3 shows that generic security solutions impose unnecessary overhead in some domains and insufficient security in others, because they adopt worst-case assumptions that do not reflect domain-specific operational constraints, threat models, and performance requirements. Security protocols designed for stationary industrial control systems are too heavyweight for resource-constrained UAV swarms, whereas protocols designed for UAV swarms cannot provide the reliability required by smart grids where availability is paramount. Domain-customized design therefore extends the reviewed methods to UAV swarms, smart grids, vehicle platoons, satellite formations, and industrial control systems[29,99,103,106,107,135,174,186,187]. For agricultural embodied intelligent robot swarms, five specific gaps follow directly from the analysis in Section 3: heterogeneous trust management between ground machinery and UAV formations, including cross-platform identity verification and semantic consistency of task-related messages; lightweight multimodal deception detection that fuses GNSS, vision, LiDAR, IMU, and

multispectral data under embedded resource constraints; resilient communication protocols for extreme farmland channels, including adaptive event-triggering, UAV relaying, and delay-tolerant networking; safe RL with physical constraints for spraying, weeding, and harvesting; and field-validated benchmarks that include realistic attack scenarios, weather conditions, crop variability, and hardware-in-the-loop testing. A dedicated agricultural security testbed with real tractors, UAVs, sprayers, and field sensors is essential to bridge simulation and deployment. The broader open questions are how to systematically map domain-specific constraints and threat models to mechanism requirements, how to certify compliance with domain regulatory standards, and how to design mechanisms that degrade gracefully under resource limits rather than failing catastrophically when the design envelope is exceeded.

4.10. Proactive defense against AI-powered adaptive attacks

Almost all defense mechanisms reviewed in Section 3 respond only after an attack is detected and assume a fixed and known attack strategy. The rapid development of artificial intelligence is giving rise to a new generation of adaptive attacks that can automatically detect vulnerabilities, optimize strategies in real time, evade detection by learning the defense mechanism, and even exploit AI-based defense components through adversarial samples and prompt injection. These attacks break the fixed attack model assumptions underlying most existing resilient control theories. As both attackers and defenders adopt AI, the security of MASs will depend on a continuous co-evolutionary arms race that requires a fundamental shift from passive response to active defense. Active deception, honeypots, watermarking, adversarially robust learning, and opponent modeling extend the detection and learning-based defenses reviewed in Sections 3.1 and 3.3 to this co-evolutionary setting[188]. The open questions are how to model the co-evolutionary dynamics between AI attackers and defenders, how to defend against prompt injection and semantic attacks on large language model integrated autonomous agents that operate at the semantic rather than network level, and how to design active defense mechanisms that avoid degrading normal system performance or introducing new vulnerabilities through the deception mechanism itself.

4.11. Summary and roadmap

Table 6 summarizes the ten directions presented in Figure 4, each linked to the open challenges identified in Section 3 and to the reviewed methods they extend.

This roadmap highlights a gradual shift from near-term standardization and domain customization, through medium-term intelligent and data-driven defense, to long-term endogenous security and formally verified MASs. Each direction is anchored in a concrete gap identified by the comparative analysis of Section 3, and progress along these ten directions will deepen both the scientific understanding of secure cooperative control and the practical deployment of autonomous MASs in critical societal infrastructure.

5. CONCLUSION

Secure control of MASs has evolved from a niche topic into a core enabling technology for autonomous cyber-physical systems. This survey has reviewed the rapidly growing literature along three complementary lines: attack modeling, defense mechanisms, and fundamental structural tensions, using the four-node information-flow architecture as the analytical backbone. Our analysis reveals a profound paradigm shift: first-generation solutions followed an add-on, design-first-then-patch model, whereas the emerging paradigm treats security as an intrinsic attribute and embeds resilience into the architecture, dynamic behavior, and learning process. This shift unfolds along three dimensions: from isolated single-point defense to collaborative swarm intelligence, from passive response to active adaptive defense, and from separate cyber and physical design to deep integration that turns physical invariants into unforgeable security primitives.

The ultimate goal is not merely to make MASs harder to compromise, but to build autonomous systems with intrinsic security, in which perception, learning, decision-making, and defense form a closed loop capable of self-monitoring, self-adaptation, and self-repair without external patches. Realizing this vision requires sustained collaboration across control, cybersecurity, machine learning, formal methods, and human-factors engineering. We have also critically examined the limitations of learning-based secure control, including the lack of stability certification, high sample complexity, unsafe exploration, and vulnerability to adversarial manipulation, underscoring that learning-based methods currently complement, rather than replace, model-based resilient control. Hybrid architectures with certified safety envelopes are a promising path forward.

Among emerging applications, agricultural embodied intelligent robot swarms are particularly notable because their safe collaboration directly affects food security and sustainable production. The proposed four-node information-flow framework can be directly instantiated in this domain: perception-layer defenses protect GNSS, visual, LiDAR, and multispectral sensing; transmission-layer defenses maintain intermittent rural connectivity; decision-layer defenses enable resilient consensus and task allocation under Byzantine or FDI-compromised neighbors; and execution-layer defenses preserve safe actuation of sprayers, weeders, and wheel motors. This node-by-node mapping provides a practical roadmap for designing, verifying, and deploying secure agricultural robot swarms in open and adversarial rural environments. The challenges and solutions reviewed in this article thus lay the foundation for a resilient agricultural robot network that can operate safely in such environments, and more broadly for a new generation of trusted autonomous systems in critical societal infrastructure.

DECLARATIONS

Authors' contributions

Conceptualization, methodology, formal analysis, writing - original draft, writing - review and editing, visualization, project administration: Huang, K.

Investigation, data curation, validation, writing - review and editing, visualization: Shi, B.

Conceptualization, methodology, supervision, resources, project administration, funding acquisition, writing - review and editing: Chen, J.

Availability of data and materials

Not applicable.

AI and AI-assisted tools statement

During preparation of this manuscript, the AI tool ChatGPT (version GPT-4, released 2023-03-14) was used solely for language editing. The tool did not influence the study design, data collection, analysis, interpretation, or the scientific content of the work. All authors take full responsibility for the accuracy, integrity, and final content of the manuscript.

Financial support and sponsorship

This work was funded by the National Natural Science Foundation of China (Grant No. 52472463), the National Key Research and Development Program of China (Grant No. 2022YFD2001405), the State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences (Grant No. 2024-O01), the Special Fund for the Construction of World-Class Universities and Disciplines of China Agricultural University (Grant No. 2025AC044), the Pinduoduo-China Agricultural University Research Fund (Grant No. PC2024A02001), the Discipline Integration and Innovation Project of China Agricultural University, and the Professor Workstation Program in Jize County of China Agricultural University, the Guizhou Provincial Key Laboratory of Mountainous Intelligent Agricultural Machinery (Grant No. 202507), and the Key Laboratory of Agricultural Equipment for Hilly and Mountainous Areas in Southeastern China (Co-construction by Ministry and Province), Ministry of Agriculture and Rural Affairs (Grant No. KFKT2025009).

Conflicts of interest

Chen, J. is a Junior Chief Editor and a Guest Editor of the special issue "Embodied Intelligence for Smart Agricultural Robotics" in the journal *Intelligence & Robotics*, but was not involved in any steps of editorial processing, notably including reviewer selection, manuscript handling, and decision making. The other authors have declared that they have no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2026.

REFERENCES

1. He, W.; Qian, F.; Lam, J.; Chen, G.; Han, Q.; Kurths, J. Quasi-synchronization of heterogeneous dynamic networks via distributed impulsive control: error estimation, optimization and design. *Automatica* **2015**, *62*, 249-62. DOI
2. Yu, X.; Zhang, Y. Sense and avoid technologies with applications to unmanned aircraft systems: review and prospects. *Prog. Aerosp. Sci.* **2015**, *74*, 152-66. DOI
3. Kang, W.; Sparks, A.; Banda, S. Coordinated control of multisatellite systems. *J. Guid. Control. Dyn.* **2001**, *24*, 360-8. DOI
4. Hausman, K.; Müller, J.; Hariharan, A.; Ayanian, N.; Sukhatme, G. S. *Cooperative control for target tracking with onboard sensing*. In Experimental robotics; Cham: Springer International Publishing; 2016. pp. 879-92. DOI
5. Ge, X.; Han, Q.; Zhong, M.; Zhang, X. Distributed Krein space-based attack detection over sensor networks under deception attacks. *Automatica* **2019**, *109*, 108557. DOI
6. Zhang, S.; Li, J.; Liu, C.; Fu, L.; Bai, Z.; Li, J. SIGMA: an agent-based modeling UAV swarm simulator for swarm intelligence algorithms. *IEEE. Trans. Autom. Sci. Eng.* **2025**, *22*, 19694-708. DOI
7. Yi, B.; Qiu, L.; Lv, J.; Nian, Y.; Wang, X.; Das, S. K. Toward fine-gained services: NFV-assisted tracking and positioning using micro-services for multi-robot cooperation. *IEEE. Netw.* **2024**, *38*, 8-16. DOI
8. Hu, T.; Wu, Z.; Feng, L. Distributed fault detection for multirate network systems over WSNs with random multiple time delays under improved low-duty-cycle scheduling. *IEEE. Sens. J.* **2026**, *26*, 1347-57. DOI
9. Guliashki, V.; Marinova, G. AI-assisted optimal energy management in microgrids. In *2025 32nd International Conference on Systems, Signals and Image Processing (IWSSIP)*, Skopje, North Macedonia. Jun 24-26, 2025. IEEE; 2025. p. 1-5. DOI
10. Abhang, L. B.; Gummadi, A.; Changala, R.; Vuyyuru, V. A.; R, S.; Raj, I. I. Swarm intelligence for multi-robot coordination in agricultural Automation. In *2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India. Mar 14-15, 2024. IEEE; 2024. pp. 455-60. DOI
11. Maitra, A. K. Offensive cyber-weapons: technical, legal, and strategic aspects. *Environ. Syst. Decis.* **2015**, *35*, 169-82. DOI
12. Xu, W.; Hu, G.; Ho, D. W. C.; Feng, Z. Distributed secure cooperative control under denial-of-service attacks from multiple adversaries. *IEEE. Trans. Cybern.* **2020**, *50*, 3458-67. DOI
13. Feng, Z.; Hu, G. Secure cooperative event-triggered control of linear multiagent systems under DoS attacks. *IEEE. Trans. Control. Syst. Technol.* **2020**, *28*, 741-52. DOI
14. Leblanc, H. J.; Zhang, H.; Koutsoukos, X.; Sundaram, S. Resilient asymptotic consensus in robust networks. *IEEE. J. Sel. Areas. Commun.* **2013**, *31*, 766-81. DOI
15. Dibaji, S. M.; Ishii, H. Consensus of second-order multi-agent systems in the presence of locally bounded faults. *Syst. Control. Lett.* **2015**, *79*, 23-9. DOI
16. Wen, G.; Yu, W.; Yu, X.; Lü, J. Complex cyber-physical networks: from cybersecurity to security control. *J. Syst. Sci. Complex.* **2017**, *30*, 46-67. DOI
17. Peng, C.; Sun, H.; Yang, M.; Wang, Y. A survey on security communication and control for smart grids under malicious cyber attacks. *IEEE. Trans. Syst. Man. Cybern. Syst.* **2019**, *49*, 1554-69. DOI
18. Zhang, W.; Wang, Z.; Liu, Y.; Ding, D.; Alsaadi, F. E. Sampled-data consensus of nonlinear multiagent systems subject to cyber attacks. *Int. J. Robust. Nonlinear. Control.* **2018**, *28*, 53-67. DOI
19. He, W.; Qian, F.; Han, Q.; Chen, G. Almost sure stability of nonlinear systems under random and impulsive sequential attacks. *IEEE. Trans. Autom. Control.* **2020**, *65*, 3879-86. DOI

20. He, W.; Xu, W.; Ge, X.; Han, Q.; Du, W.; Qian, F. Secure control of multiagent systems against malicious attacks: a brief survey. *IEEE. Trans. Ind. Inform.* **2022**, *18*, 3595-608. DOI
21. Ding, D.; Han, Q.; Ge, X.; Wang, J. Secure state estimation and control of cyber-physical systems: a survey. *IEEE. Trans. Syst. Man. Cybern. Syst.* **2021**, *51*, 176-90. DOI
22. Feng, Z.; Wen, G.; Hu, G. Distributed secure coordinated control for multiagent systems under strategic attacks. *IEEE. Trans. Cybern.* **2017**, *47*, 1273-84. DOI PubMed
23. Teixeira, A.; Shames, I.; Sandberg, H.; Johansson, K. H. A secure control framework for resource-limited adversaries. *Automatica* **2015**, *51*, 135-48. DOI
24. De Persis, C.; Tesi, P. Input-to-state stabilizing control under denial-of-service. *IEEE. Trans. Autom. Control.* **2015**, *60*, 2930-44. DOI
25. Li, X. M.; Yao, D.; Li, P.; Meng, W.; Li, H.; Lu, R. Secure finite-horizon consensus control of multiagent systems against cyber attacks. *IEEE. Trans. Cybern.* **2022**, *52*, 9230-9. DOI
26. Li, X.; Wang, X.; Yu, D. Impulsive synchronization control for multiagent systems under hybrid attacks. *IEEE. Syst. J.* **2025**, *19*, 883-94. DOI
27. Sun, Y.; Yang, G. Distributed dynamic event-triggered strategy for the consensus of multi-agent systems under asynchronous denial-of-service attacks. *J. Franklin. Inst.* **2026**, *363*, 108224. DOI
28. Wan, M.; Xu, Y.; Wang, L.; Wu, Y.; Wu, Z. Multi-channel asynchronous DoS-resilient control for uncertain MASs using an equivalent decay-rate approach. *IEEE. Trans. Autom. Sci. Eng.* **2026**, *23*, 7354-65. DOI
29. Deng, C.; Wen, C. MAS-based distributed resilient control for a class of cyber-physical systems with communication delays under DoS attacks. *IEEE. Trans. Cybern.* **2021**, *51*, 2347-58. DOI
30. Hong, Y.; Su, Y.; Cai, H. A resilient adaptive event-triggered distributed observer for multiagent systems under DoS attacks and its applications. *IEEE. Syst. J.* **2026**, *20*, 436-47. DOI
31. Xu, Y.; Liang, H.; Li, T.; Long, Y.; Cheng, Y.; Wang, D. Adaptive fuzzy resilient control of nonlinear multiagent systems under DoS attacks: a dynamic event-triggered method. *IEEE. Trans. Fuzzy. Syst.* **2024**, *32*, 3568-80. DOI
32. Zhang, T.; Ye, D.; Guo, G. Distributed event-triggered control for multiagent systems under denial-of-service attacked topology: secure mode strategy. *IEEE. Trans. Syst. Man. Cybern. Syst.* **2022**, *52*, 6534-44. DOI
33. Lin, X.; Zhang, Z.; Ren, L.; Sun, M.; Liu, J.; Sun, C. Intermittent event-triggered fixed-time containment control for disturbed second-order multi-agent systems under DoS attacks. *IEEE. Trans. Signal. Inf. Process. Netw.* **2026**, *12*, 756-70. DOI
34. Dai, H.; Ji, L.; Guo, X.; Zhang, C.; Yang, S.; Li, H. Leader-following synchronization control of multiagent systems under hybrid cyber attacks via impulsive control based on topology switching. *IEEE. Trans. Cybern.* **2024**, *54*, 5297-308. DOI
35. Deng, C.; Wen, C. Distributed resilient observer-based fault-tolerant control for heterogeneous multiagent systems under actuator faults and DoS attacks. *IEEE. Trans. Control. Netw. Syst.* **2020**, *7*, 1308-18. DOI
36. Zhang, Z.; Dong, J. Fault-tolerant containment control for IT2 fuzzy networked multiagent systems against denial-of-service attacks and actuator faults. *IEEE. Trans. Syst. Man. Cybern. Syst.* **2022**, *52*, 2213-24. DOI
37. Jiang, Y.; Niu, B.; Zhao, T.; Zhao, X.; Wang, X.; Wang, H. Intelligent consensus asymptotic tracking control for nonlinear multiagent systems under denial-of-service attacks. *IEEE. Trans. Autom. Sci. Eng.* **2025**, *22*, 776-87. DOI
38. Xu, Y.; Wu, Z. Data-based collaborative learning for multiagent systems under distributed denial-of-service attacks. *IEEE. Trans. Cogn. Dev. Syst.* **2024**, *16*, 75-85. DOI
39. Ni, J.; Duan, F.; Shi, P. Fixed-time consensus tracking of multiagent system under DOS attack with event-triggered mechanism. *IEEE. Trans. Circuits. Syst. I. Regul. Pap.* **2022**, *69*, 5286-99. DOI
40. Li, R.; Gan, Q.; Wu, H.; Cao, J.; Kang, Q. Resilient finite-time consensus of variable-order fractional multiagent systems under DoS attacks. *IEEE. Trans. Syst. Man. Cybern. Syst.* **2025**, *55*, 3188-201. DOI
41. Yang, Y.; Bi, S.; Dai, R.; Shen, Q. Self-triggered predefined-time cooperative control against DoS attacks for multiagent systems with uncertain powers. *IEEE. Trans. Cybern.* **2025**, *55*, 5202-12. DOI PubMed
42. Nian, X.; Li, F.; Liu, D. Distributed algorithm for constrained social cost minimization problem of heterogeneous linear multiagent systems against DoS attacks. *IEEE. Trans. Control. Netw. Syst.* **2024**, *11*, 2101-13. DOI
43. Cui, Y.; Jenabzadeh, A.; Ahmed, Z.; Zhang, W.; Huang, T. Resilient distributed control and target tracking in multiagent systems against composite attacks. *IEEE. Trans. Syst. Man. Cybern. Syst.* **2025**, *55*, 1252-63. DOI
44. Gil, S.; Yemini, M.; Chorti, A.; Nedić, A.; Poor, H. V.; Goldsmith, A. J. How physicality enables Cy-trust: a new era of trust-centered cyber–physical systems. *Proc. IEEE.* **2025**, *113*, 1121-54. DOI
45. Finistrella, S.; Mariani, S.; Zambonelli, F. Multi-agent reinforcement learning for cybersecurity: classification and survey. *Intell. Syst. Appl.* **2025**, *26*, 200495. DOI
46. Giraldo, J.; Urbina, D.; Cardenas, A.; et al. A survey of physics-based attack detection in cyber-physical systems. *ACM. Comput. Surv.* **2018**, *51*, 1-36. DOI PubMed PMC

47. Ding, D.; Han, Q.; Xiang, Y.; Ge, X.; Zhang, X. A survey on security control and attack detection for industrial cyber-physical systems. *Neurocomputing* **2018**, *275*, 1674-83. DOI
48. Dibaji, S. M.; Pirani, M.; Flamholz, D. B.; Annaswamy, A. M.; Johansson, K. H.; Chakrabortty, A. A systems and control perspective of CPS security. *Annu. Rev. Control.* **2019**, *47*, 394-411. DOI
49. Cetinkaya, A.; Ishii, H.; Hayakawa, T. An overview on denial-of-service attacks in control systems: attack models and security analyses. *Entropy* **2019**, *21*, 210. DOI PubMed PMC
50. Olfati-Saber, R.; Murray, R. Consensus problems in networks of agents with switching topology and time-delays. *IEEE. Trans. Autom. Control.* **2004**, *49*, 1520-33. DOI
51. Ren, W.; Beard, R. W. *Distributed consensus in multi-vehicle cooperative control*; 1st ed. London: Springer London; 2008. DOI
52. Amin, S.; Cárdenas, A. A.; Sastry, S. S. *Safe and secure networked control systems under denial-of-service attacks*. In Hybrid systems: computation and control. Berlin: Springer Berlin Heidelberg; 2009. pp. 31-45. DOI
53. Pasqualetti, F.; Dorfler, F.; Bullo, F. Attack detection and identification in cyber-physical systems. *IEEE. Trans. Autom. Control.* **2013**, *58*, 2715-29. DOI
54. Dolk, V. S.; Borgers, D. P.; Heemels, W. P. M. H. Output-based and decentralized dynamic event-triggered control with guaranteed $\mathcal{L}_p$-gain performance and zeno-freeness. *IEEE. Trans. Autom. Control.* **2017**, *62*, 34-49. DOI
55. Zhu, H.; Li, Y.; Tong, S. Dynamic event-triggered reinforcement learning control of stochastic nonlinear systems. *IEEE. Trans. Fuzzy. Syst.* **2023**, *31*, 2917-28. DOI
56. Cui, Y.; Cao, L.; Gong, X.; Basin, M. V.; Shen, J.; Huang, T. Resilient output containment control of heterogeneous multiagent systems against composite attacks: a digital twin approach. *IEEE. Trans. Cybern.* **2024**, *54*, 3313-26. DOI
57. He, J.; Guo, X.; Chen, Z.; Kuang, Y. Secure impulsive synchronization control of multi-agent systems under switching deception attacks on dual channel. *Commun. Nonlinear. Sci. Numer. Simul.* **2025**, *149*, 108922. DOI
58. He, W.; Mo, Z. Secure event-triggered consensus control of linear multiagent systems subject to sequential scaling attacks. *IEEE. Trans. Cybern.* **2022**, *52*, 10314-27. DOI PubMed
59. Shang, Y. Resilient vector consensus over random dynamic networks under mobile malicious attacks. *Comput. J.* **2024**, *67*, 1076-86. DOI
60. Wang, Y.; Fang, Z.; Wang, J.; Xiong, J.; Feng, S.; Zhang, G. Architecture-driven robotic cybersecurity: a systematic review and taxonomy. *Comput. Netw.* **2026**, *287*, 112590. DOI
61. Chhabra, A.; Datta, S.; Nahin, S. K.; Mohapatra, P. Agentic AI security: threats, defenses, evaluation, and open challenges. *IEEE. Access.* **2026**, *14*, 49455-82. DOI
62. Wang, X.; Garg, S.; Lin, H.; et al. Toward accurate anomaly detection in Industrial Internet of Things using hierarchical federated learning. *IEEE. Internet. Things. J.* **2022**, *9*, 7110-9. DOI
63. Zhu, H.; Liu, J.; Zhang, Z.; Zhang, S.; Qu, F. Consensus control of multiagent systems under switching topology and denial-of-service attacks. *IEEE. Syst. J.* **2023**, *17*, 5911-9. DOI
64. Godsil, C.; Royle, G. *Algebraic graph theory*; 1st ed. Springer New York, NY; 2001. DOI
65. Mesbahi, M.; Egerstedt, M. *Graph theoretic methods in multiagent networks*. Princeton: Princeton University Press; 2010. DOI
66. Li, Z.; Duan, Z.; Chen, G.; Huang, L. Consensus of multiagent systems and synchronization of complex networks: a unified viewpoint. *IEEE. Trans. Circuits. Syst. I. Regul. Pap.* **2010**, *57*, 213-24. DOI
67. Jadbabaie, A.; Lin, J.; Morse, A. S. Coordination of groups of mobile autonomous agents using nearest neighbor rules. *IEEE. Trans. Autom. Control.* **2003**, *48*, 988-1001. DOI
68. Wieland, P.; Sepulchre, R.; Allgöwer, F. An internal model principle is necessary and sufficient for linear output synchronization. *Automatica* **2011**, *47*, 1068-74. DOI
69. Krstic, M.; Kanellakopoulos, I.; Kokotović, P. V. *Nonlinear and adaptive control design*. New York: John Wiley & Sons; 1995. https://www.wiley.com/en-us/shop/general-introductory-electrical-electronics-engineering/nonlinear-and-adaptive-control-design-p-9780471127321. (accessed 2026-11-10).
70. Slotine, J. J. E.; Li, W. P. *Applied nonlinear control*. China Machine Press; 1991. https://www.matlabi.ir/wp-content/uploads/bank_papers/g_paper/g465-www.matlabi.ir.pdf. (accessed 2026-11-04).
71. Robot modeling and control, second edition [Bookshelf]. *IEEE. Control. Syst.* **2022**, *42*, 126-8. DOI
72. Kelly, R.; Davila, V. S.; Perez, J. A. L. *Control of robot manipulators in joint space*; 1st ed. Springer London; 2005. DOI
73. Jin, B. *Fractional differential equations: an approach via fractional derivatives*; 1st ed. Springer Cham; 2021. DOI
74. Khasminskii, R. *Stochastic stability of differential equations*; 2nd ed. Springer Berlin, Heidelberg; 2012. DOI
75. Deng, H.; Krstić, M. Stochastic nonlinear stabilization - I: A backstepping design. *Syst. Control. Lett.* **1997**, *32*, 143-50. DOI
76. Zhang, Y.; Jiang, J. Bibliographical review on reconfigurable fault-tolerant control systems. *Annu. Rev. Control.* **2008**, *32*, 229-52. DOI

77. Polycarpou, M. M.; Helmicki, A. J. Automated fault detection and accommodation: a learning systems approach. *IEEE. Trans. Syst. Man. Cybern.* **1995**, *25*, 1447-58. DOI
78. Hu, T.; Lin, Z. *Control systems with actuator saturation: analysis and design*; 1st ed. Boston: Birkhäuser; 2001. DOI
79. Tarbouriech, S.; Garcia, G.; da Silva, J. M. G. Jr.; Queinnec, I. *Stability and stabilization of linear systems with saturating actuators*; 1st ed. London: Springer London; 2011. DOI
80. Yan, B.; Shi, P.; Chambers, J. Cooperative control for heterogeneous multi-agent systems: progress, applications, and challenges. *Sci. China. Inf. Sci.* **2024**, *67*, 4001. DOI
81. Feng, Z.; Hu, G.; Wen, G. Distributed consensus tracking for multi-agent systems under two types of attacks. *Int. J. Robust. Nonlinear. Control.* **2016**, *26*, 896-918. DOI
82. Pasha, S. A.; Safdar, R.; Ali, S. T. False data injection attacks on networked control systems. *J. Control. Decis.* **2024**, *11*, 650-9. DOI
83. Mo, Y.; Sinopoli, B. Secure control against replay attacks. In *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Monticello, USA. Sep 30 - Oct 02, 2009. IEEE; 2009. pp. 911-8. DOI
84. Lynch, N. A. *Distributed algorithms*. Morgan Kaufmann Publishers Inc.; 1996. DOI
85. Sundaram, S.; Gharesifard, B. Distributed optimization under adversarial nodes. *IEEE. Trans. Autom. Control.* **2019**, *64*, 1063-76. DOI
86. Oh, K.; Park, M.; Ahn, H. A survey of multi-agent formation control. *Automatica* **2015**, *53*, 424-40. DOI
87. Zhao, S.; Zelazo, D. Bearing rigidity theory and its applications for control and estimation of network systems: life beyond distance rigidity. *IEEE. Control. Syst.* **2019**, *39*, 66-83. DOI
88. Ji, M.; Ferrari-Trecate, G.; Egerstedt, M.; Buffa, A. Containment control in mobile networks. *IEEE. Trans. Autom. Control.* **2008**, *53*, 1972-5. DOI
89. Cao, Y.; Ren, W.; Egerstedt, M. Distributed containment control with multiple stationary or dynamic leaders in fixed and switching directed networks. *Automatica* **2012**, *48*, 1586-97. DOI
90. Francis, B. A. The linear multivariable regulator problem. In *1976 IEEE Conference on Decision and Control including the 15th Symposium on Adaptive Processes*, Clearwater, USA. Dec 01-03, 1976. IEEE; 1976. pp. 873-8. DOI
91. Huang, J. *Nonlinear output regulation: theory and applications*. 2004. DOI
92. Nedic, A.; Ozdaglar, A. Distributed subgradient methods for multi-agent optimization. *IEEE. Trans. Autom. Control.* **2009**, *54*, 48-61. DOI
93. Wang, J.; Elia, N. A control perspective for centralized and distributed convex optimization. In *2011 50th IEEE Conference on Decision and Control and European Control Conference*, Orlando, USA. Dec 12-15, 2011. IEEE; 2011. pp. 3800-5. DOI
94. Cai, J.; Wen, L.; Feng, H.; et al. An overview of security threats, attack detection and defense for large-scale multi-agent systems (LSMAS) in Internet of Things (IoT). *IEEE. Trans. Ind. Cyber. Phys. Syst.* **2025**, *3*, 70-81. DOI
95. Dkhil, R.; Boussaid, B.; Zouinkhi, A.; Djemai, M. Cyberattack detection in multi-agent system using luenberger observer under FDI attack. In *2025 IEEE International Multi-Conference on Smart Systems & Green Process (IMC-SSGP)*, Hammamet, Tunisia. Oct 30 - Nov 02, 2025. IEEE; 2025. p. 1-6. DOI
96. Zhou, S.; Liu, C.; Zhao, J.; Xu, L. Complementary dual observer-based distinguishable detection of heterogeneous multi-agent systems. In *2025 37th Chinese Control and Decision Conference (CCDC)*, Xiamen, China. May 16-19, 2025. IEEE; 2025. pp. 1160-5. DOI
97. Xu, Y.; Fang, M.; Shi, P.; Wu, Z. G. Event-based secure consensus of mutiagent systems against DoS attacks. *IEEE. Trans. Cybern.* **2020**, *50*, 3468-76. DOI
98. Yu, G.; Liu, D.; Li, L.; Jing, Y.; Shen, M. A dynamic event-triggered resilient control design for linear multi-agent systems under DoS attacks. *Asian. J. Control.* **2025**. DOI
99. Li, B.; Wang, Z.; Wang, W.; Gao, Q.; Lü, J. Event-/self-triggered communication for DoS-resilient consensus in multiagent systems with application to LEO satellite formation. *IEEE. Trans. Ind. Inform.* **2026**, *22*, 5049-60. DOI
100. Ren, Z.; Liu, H.; Basin, M. V.; Gao, Q. Event-triggered data-driven formation control for heterogeneous multiagent systems subject to DoS attacks. *IEEE. Trans. Control. Netw. Syst.* **2025**, *12*, 1968-79. DOI
101. Zhang, H.; Huang, Y.; Zhao, N.; Qiu, X.; Herrera-Viedma, E.; Agarwal, R. K. Resilient consensus control of nonlinear multiagent systems under hybrid cyberattacks: a disturbance observer-based neural network approach. *IEEE. Trans. Cybern.* **2026**, *56*, 4108-19. DOI
102. Zhang, D.; Deng, C.; Feng, G. Resilient cooperative output regulation for nonlinear multiagent systems under DoS attacks. *IEEE. Trans. Autom. Control.* **2023**, *68*, 2521-8. DOI
103. Wu, H. Y.; Liu, Q.; Park, J. H. Observer-based secure consensus for multiagent systems under multimode DoS attacks with application to power system. *IEEE. Trans. Cybern.* **2025**, *55*, 1032-44. DOI PubMed
104. Tan, M.; Liu, Z.; Chen, C.; Wang, Y.; Chen, C. L. P. Adaptive optimal consensus control for nonlinear uncertain multiagent systems under DoS attacks. *IEEE. Trans. Cybern.* **2025**, *55*, 5480-91. DOI

105. Du, Z.; Kao, Y.; Liu, Y.; Wu, Z. Dynamic event-triggered optimal control for fuzzy multiagent systems with DoS attacks. *IEEE. Trans. Autom. Sci. Eng.* **2025**, *22*, 19814-23. DOI
106. Xu, Z.; Si, C.; Liu, Y.; Wang, K.; Mu, C. Resilient control of multi agent systems against actuator attack and applications to connected vehicle system. In *2026 6th international conference on neural networks, information and communication engineering (NNICE)*, Hefei, China. Jan 23-25, 2026. IEEE; 2026. pp. 1324-9. DOI
107. Nallagatla, R.; Srivas, T.; Kumar, S.; Singh, P. Containment control of quadrotors in multi-agent system framework under communication faults. In *2025 7th International Conference on Electrical, Control and Instrumentation Engineering (ICECIE)*, Pattaya City, Thailand. Nov 22-23, 2025. IEEE; 2025. pp. 434-9. DOI
108. Li, K.; Song, L.; Li, Y. Fuzzy adaptive resilient output control for nonlinear multi-agent systems under Byzantine agents. *IEEE. Trans. Autom. Sci. Eng.* **2025**, *22*, 8783-93. DOI
109. Huang, Z.; Li, T.; Long, Y.; Shao, J. Human-in-the-loop prescribed performance optimal control for multiagent systems under DoS attacks: an error reconstruction mechanism. In *2025 17th International Conference on Advanced Computational Intelligence (ICACI)*, Bath, United Kingdom. Jul 07-13, 2025. IEEE; 2025. pp. 9-17. DOI
110. Dong, Y.; Cui, X.; Hong, Y.; Chen, J. Robust safety-critical control of uncertain nonlinear systems with DoS attacks. *IEEE. Trans. Autom. Control.* **2026**, *71*, 871-85. DOI
111. Fawzi, H.; Tabuada, P.; Diggavi, S. Secure estimation and control for cyber-physical systems under adversarial attacks. *IEEE. Trans. Autom. Control.* **2014**, *59*, 1454-67. DOI
112. Wen, G.; Wang, P.; Lv, Y.; Chen, G.; Zhou, J. Secure consensus of multi-agent systems under denial-of-service attacks. *Asian. J. Control.* **2023**, *25*, 695-709. DOI
113. Yuan, Y.; Yuan, H.; Guo, L.; Yang, H.; Sun, S. Resilient control of networked control system under DoS attacks: a unified game approach. *IEEE. Trans. Ind. Inform.* **2016**, *12*, 1786-94. DOI
114. Blanke, M.; Kinnaert, M.; Lunze, J.; Staroswiecki, M. *Diagnosis and fault-tolerant control*; 2nd ed. Springer Berlin, Heidelberg; 2006. https://books.google.com/books/about/Diagnosis_and_Fault_Tolerant_Control.html?id=FeRDAAAAQBAJ. (accessed 2026-11-04).
115. Cheng, P.; Ye, S.; Liu, R.; He, S.; Zhang, W. SMC-based bounded consensus tracking for multiagent systems under stochastic DoS attacks with applications to multiple DC motors. *IEEE. Trans. Ind. Inform.* **2024**, *20*, 10074-83. DOI
116. Nussbaum, R. D. Some remarks on a conjecture in parameter adaptive control. *Syst. Control. Lett.* **1983**, *3*, 243-6. DOI
117. Shoukry, Y.; Nuzzo, P.; Sangiovanni-Vincentelli, A. L.; Seshia, S. A.; Pappas, G. J.; Tabuada, P. SMC: satisfiability modulo convex programming. *Proc. IEEE.* **2018**, *106*, 1655-79. DOI
118. Pajic, M.; Lee, I.; Pappas, G. J. Attack-resilient state estimation for noisy dynamical systems. *IEEE. Trans. Control. Netw. Syst.* **2017**, *4*, 82-92. DOI
119. Guo, Z.; Wei, Q.; Zhao, X.; Wang, B.; Niu, B.; Liu, H. Attack detection and active attack defense for cyber - physical systems via zonotopic observer and reachability analysis. *IEEE. Trans. Cybern.* **2026**, *56*, 4586-96. DOI
120. Ma, Y. S.; Che, W. W.; Deng, C.; Wu, Z. G. Observer-based event-triggered containment control for MASs under DoS attacks. *IEEE. Trans. Cybern.* **2022**, *52*, 13156-67. DOI
121. Zhang, Y.; Wu, Z.; Shi, P.; Huang, T.; Chakrabarti, P. Quantization-based event-triggered consensus of multiagent systems against aperiodic DoS attacks. *IEEE. Trans. Syst. Man. Cybern. Syst.* **2023**, *53*, 3774-83. DOI
122. Yang, G.; Sun, Z.; Sircar, A.; et al. CuSum-based change detection of false data injection attacks in DC microgrids. *IEEE. J. Emerg. Sel. Top. Ind. Electron.* **2026**, *7*, 1585-97. DOI
123. Combastel, C. An extended zonotopic and Gaussian Kalman filter (EZGKF) merging set-membership and stochastic paradigms: toward non-linear filtering and fault detection. *Annu. Rev. Control.* **2016**, *42*, 232-43. DOI
124. Xiao, L.; Li, Y.; Han, G.; Liu, G.; Zhuang, W. PHY-layer spoofing detection with reinforcement learning in wireless networks. *IEEE. Trans. Veh. Technol.* **2016**, *65*, 10037-47. DOI
125. Zhu, Q.; Basar, T. Game-theoretic methods for robustness, security, and resilience of cyberphysical control systems: games-in-games principle for optimal cross-layer resilient control systems. *IEEE. Control. Syst. Mag.* **2015**, *35*, 46-65. DOI
126. Tabuada, P. Event-triggered real-time scheduling of stabilizing control tasks. *IEEE. Trans. Autom. Control.* **2007**, *52*, 1680-5. DOI
127. Heemels, W. P. M. H.; Johansson, K. H.; Tabuada, P. An introduction to event-triggered and self-triggered control. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, Maui, USA. Dec 10-13, 2012. IEEE; 2012. pp. 3270-85. DOI
128. Xu, Y.; Zhao, H.; Gao, Y.; Yu, H.; Peng, L. Data-based dynamic event-triggered attitude control for helicopter with aperiodic denial-of-service attacks. *Intell. Robot.* **2025**, *5*, 764-82. DOI
129. Wang, W.; Nešić, D.; Postoyan, R.; Heemels, W. P. M. H. A unifying event-triggered control framework based on a hybrid small-gain theorem. In *2020 59th IEEE Conference on Decision and Control (CDC)*, Jeju, Korea. Dec 14-18, 2020. IEEE; 2020. pp. 4979-84. DOI
130. Yi, X.; Liu, K.; Dimarogonas, D. V.; Johansson, K. H. Dynamic event-triggered and self-triggered control for multi-agent systems. *IEEE. Trans. Autom. Control.* **2019**, *64*, 3300-7. DOI

131. Zhou, S.; Liu, Y.; Lu, X.; Li, W.; Long, J. Dynamic event-triggered cluster consensus for multiagent systems under DoS attacks with antagonistic interactions. *IEEE. Trans. Syst. Man. Cybern. Syst.* **2025**, *55*, 4362-74. DOI
132. De Persis, C.; Tesi, P. Networked control of nonlinear systems under denial-of-service. *Syst. Control. Lett.* **2016**, *96*, 124-31. DOI
133. Mazo, M.; Anta, A.; Tabuada, P. An ISS self-triggered implementation of linear controllers. *Automatica* **2010**, *46*, 1310-4. DOI
134. Anta, A.; Tabuada, P. To sample or not to sample: self-triggered control for nonlinear systems. *IEEE. Trans. Autom. Control.* **2010**, *55*, 2030-42. DOI
135. Liang, H.; Chang, Z.; Pan, Y. Dual-event-triggered intelligence security control for multiagent systems against DoS attacks with applications in mobile robot systems. *IEEE. Trans. Artif. Intell.* **2024**, *5*, 916-24. DOI
136. Sun, W.; Liu, S.; Yang, D.; Zhang, J. Formation tracking control of multi-agent systems with non-zero leader input under DoS attacks by asynchronous self-triggered mechanism. *J. Autom. Intell.* **2025**, *In Press*. DOI
137. Liberzon, D. *Switching in systems and control*; 1st ed. Birkhäuser Boston, MA; 2003. DOI
138. Ceragioli, F.; De Persis, C.; Frasca, P. Discontinuities and hysteresis in quantized average consensus. *Automatica* **2011**, *47*, 1916-28. DOI
139. Nesic, D.; Liberzon, D. A unified framework for design and analysis of networked and quantized control systems. *IEEE. Trans. Autom. Control.* **2009**, *54*, 732-47. DOI
140. Ling, Q.; Lemmon, M. D. Stability of quantized control systems under dynamic bit assignment. *IEEE. Trans. Autom. Control.* **2005**, *50*, 734-40. DOI
141. Hespanha, J. P.; Morse, A. S. Stability of switched systems with average dwell-time. In *Proceedings of the 38th IEEE Conference on Decision and Control (Cat. No.99CH36304)*, Phoenix, USA. Dec 07-10, 1999. IEEE; 1999. pp. 2655-60. DOI
142. Liberzon, D.; Morse, A. S. Basic problems in stability and design of switched systems. *IEEE. Control. Syst. Mag.* **1999**, *19*, 59-70. DOI
143. Li, Y.; Xiao, L.; Dai, H.; Poor, H. V. Game theoretic study of protecting MIMO transmissions against smart attacks. In *2017 IEEE International Conference on Communications (ICC)*, Paris, France. May 21-25, 2017. IEEE; 2017. p. 1-6. DOI
144. Narendra, K. S.; Annaswamy, A. M. Robust adaptive control. In *1984 American Control Conference*, San Diego, USA. Jun 06-08, 1984. IEEE; 1984. pp. 333-5. DOI
145. Hao, C.; Liu, Z.; Zheng, L.; Philip Chen, C.; Lai, G. Observer-based distributed adaptive neural network containment control for uncertain nonlinear multi-agent systems under DoS attacks. *Neurocomputing* **2026**, *659*, 131306. DOI
146. Wang, L. X. *A course in fuzzy systems and control*. 1996. https://api.semanticscholar.org/CorpusID:60936449. (accessed 2026-11-04).
147. Ge, S. S.; Hang, C. C.; Lee, T. H.; Zhang, T. *Stable adaptive neural network control*; 1st ed. Springer New York, NY; 2002. DOI
148. Bu, X.; Guo, J.; Cui, L.; Hou, Z. Event-triggered model-free adaptive containment control for nonlinear multiagent systems under DoS attacks. *IEEE. Trans. Control. Netw. Syst.* **2024**, *11*, 1845-57. DOI
149. Lewis, F. L.; Vrabie, D.; Vamvoudakis, K. G. Reinforcement learning and feedback control: using natural decision methods to design optimal adaptive controllers. *IEEE. Control. Syst. Mag.* **2012**, *32*, 76-105. DOI
150. Vamvoudakis, K. G.; Antsaklis, P. J.; Dixon, W. E.; et al. Autonomy and machine intelligence in complex systems: a tutorial. In *2015 American Control Conference (ACC)*, Chicago, USA. Jul 01-03, 2015. IEEE; 2015. pp. 5062-79. DOI
151. Modares, H.; Lewis, F. L.; Naghibi-Sistani, M. Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems. *Automatica* **2014**, *50*, 193-202. DOI
152. Yu, C.; Velu, A.; Vinitsky, E.; Gao, J.; Wang, Y.; et al. The surprising effectiveness of PPO in cooperative multi-agent games. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*. 2022. Curran Associates, Inc.; 2022. pp. 24611-24. DOI
153. Zuo, Z.; Han, Q.; Ning, B.; Ge, X.; Zhang, X. An overview of recent advances in fixed-time cooperative control of multiagent systems. *IEEE. Trans. Ind. Inform.* **2018**, *14*, 2322-34. DOI
154. Lu, Z.; Liu, G.; Lai, L.; Xu, W. Camouflage adversarial attacks on multi-agent reinforcement learning systems. *IEEE. Trans. Signal. Process.* **2026**, *74*, 589-604. DOI
155. Sudheer, A.; Chou, C. H.; Kumar, S. Demo: a real-time multi-agent network attack detection and incident response system. In *2025 Silicon Valley Cybersecurity Conference (SVCC)*, San Francisco, USA. Jun 23-25, 2025. IEEE; 2025. p. 1-3. DOI
156. An, X.; Zhang, C. LSTM-based distributed attacks detection in mass: a framework comparison study. In *2026 5th Conference on Fully Actuated System Theory and Applications (FASTA)*, Qinhuangdao, China. May 22-24, 2026. IEEE; 2026. pp. 1435-40. DOI
157. Zha, L.; Wang, J.; Cao, J. Differential privacy-based resilient bipartite consensus for multi-agent systems under adversarial nodes. *IEEE. Trans. Consum. Electron.* **2026**, *72*, 6447-55. DOI
158. Zhao, W.; Ni, Z.; Liu, J.; Zhang, H.; Yu, W. Differentially private model-free adaptive consensus control for nonlinear multiagent systems under FDI attacks. *IEEE. Internet. Things. J.* **2026**, *13*, 29200-10. DOI
159. Doyle, J. Robust and optimal control. In *Proceedings of 35th IEEE Conference on Decision and Control*, Kobe, Japan. Dec 13-13, 1996. IEEE; 1996. pp. 1595-8. DOI

160. Wu, X.; Ding, S.; Zhao, N.; Wang, H.; Niu, B. Neural-network-based event-triggered adaptive secure fault-tolerant containment control for nonlinear multi-agent systems under denial-of-service attacks. *Neural. Netw.* **2025**, *191*, 107725. DOI
161. Utkin, V. I. *Sliding modes in control and optimization*; 1st ed. Springer Berlin, Heidelberg; 1992. DOI
162. Edwards, C.; Spurgeon, S. K. Sliding mode control: theory and applications. CRC Press; 1998. DOI
163. Fu, Y.; Wen, G.; Zhao, D.; Zheng, W. X.; Li, X. Event-triggered resilient consensus of networked euler–lagrange systems under Byzantine attacks. *IEEE. Trans. Autom. Control.* **2025**, *70*, 8282-9. DOI
164. Chen, X.; Hu, S.; Yang, T.; Xie, X.; Qiu, J. Event-triggered bipartite consensus of multiagent systems with input saturation and DoS attacks over weighted directed networks. *IEEE. Trans. Syst. Man. Cybern. Syst.* **2024**, *54*, 4054-65. DOI
165. Yang, T. Impulsive control. *IEEE. Trans. Autom. Control.* **1999**, *44*, 1081-3. DOI
166. Li, X.; Song, S. Stabilization of delay systems: delay-dependent impulsive control. *IEEE. Trans. Autom. Control.* **2017**, *62*, 406-11. DOI
167. Nozari, E.; Tallapragada, P.; Cortés, J. Differentially private average consensus: obstructions, trade-offs, and optimal algorithm design. *Automatica* **2017**, *81*, 221-31. DOI
168. Kim, J.; Lee, C.; Shim, H.; et al. Encrypting controller using fully homomorphic encryption for security of cyber-physical systems. *IFAC-PapersOnLine* **2016**, *49*, 175-80. DOI
169. Piran, F. J.; Chen, Z.; Imani, M.; Imani, F. Privacy-preserving federated learning with differentially private hyperdimensional computing. *Comput. Electr. Eng.* **2025**, *123*, 110261. DOI
170. Wang, Y. Privacy-preserving average consensus via state decomposition. *IEEE. Trans. Autom. Control.* **2019**, *64*, 4711-6. DOI
171. Huo, X.; Liu, M. Privacy-preserving distributed multi-agent cooperative optimization - paradigm design and privacy analysis. *IEEE. Control. Syst. Lett.* **2022**, *6*, 824-9. DOI
172. Xu, C.; Liu, Q. A privacy-preserving trust-based resilient algorithm for leader-follower consensus against external attacks. In *2025 Joint International Conference on Automation-Intelligence-Safety (ICAIS) & International Symposium on Autonomous Systems (ISAS)*, Xi'an, China. May 23-25, 2025. IEEE; 2025. p. 1-5. DOI
173. Wang, Y.; Rajabinezhad, M.; Zhang, Y.; Zuo, S. Cyber-physical defense for heterogeneous multi-agent systems against exponentially unbounded attacks on signed digraphs. *IEEE. Trans. Netw. Sci. Eng.* **2025**, *12*, 2165-79. DOI
174. Mokari, H.; Firouzmand, E.; Sharifi, I.; Doustmohammadi, A. DoS attack detection and resilient control in platoon of smart vehicles. In *2021 9th RSI International Conference on Robotics and Mechatronics (ICRoM)*, Tehran, Iran. Nov 17-19, 2021. IEEE; 2021. pp. 144-50. DOI
175. Su, L.; Hua, Y.; Dong, X.; Lü, J.; Wang, D. Distributed resilient time-varying affine formation optimal tracking for multiagent systems against Byzantine attacks. *IEEE. Trans. Control. Netw. Syst.* **2026**, *13*, 462-74. DOI
176. Zhang, Y.; Ma, L.; Gao, C. Consensus control of multiagent systems under DoS attacks: a dynamic-key-based secure scheme. *IEEE. Trans. Cybern.* **2026**, *56*, 4417-29. DOI PubMed
177. Yang, C.; Che, W. Data-driven attack detection and resilient control for heterogeneous nonlinear nonaffine MASs. *IEEE. Trans. Netw. Sci. Eng.* **2025**, *12*, 3812-23. DOI
178. Xiong, H.; Chen, G.; Ren, H.; Li, H. Broad-learning-system-based model-free adaptive predictive control for nonlinear MASs under DoS attacks. *IEEE/CAA. J. Autom. Sin.* **2025**, *12*, 381-93. DOI
179. Zhang, L.; Zhang, R.; Park, J. H.; Chen, M.; Zeng, D. A new logical security protection control protocol for secure consensus of MASs under mixed attacks. *IEEE. Internet. Things. J.* **2026**, *13*, 18807-21. DOI
180. Liu, R.; Li, T.; Hu, S.; Mangini, A. M.; Fanti, M. P. Securing networked discrete event systems for diagnosability under attacks. *IEEE. Trans. Autom. Sci. Eng.* **2026**, *23*, 7409-23. DOI
181. Rehan, M.; Rehman, A. U.; Ahmed, I.; Khalid, M.; Ahn, C. K. A multiobjective output-feedback framework for fully distributed resilient consensus of multiagents under stochastic DoS-attacks. *IEEE. Internet. Things. J.* **2026**, *13*, 27493-504. DOI
182. Hussain, M.; Tahir, F.; Rehan, M.; Ahn, C. K.; Ahmed, I.; Khalid, M. Fully-distributed resilient consensus of multi-agent systems with state and input sector nonlinearities under stochastic denial-of-service attacks. *IEEE. Trans. Autom. Sci. Eng.* **2025**, *22*, 19582-93. DOI
183. Xu, Q.; Guo, G. Performance-guaranteed low-complexity control of multiagent systems under DoS attacks: a relaxed virtual condition design. *IEEE. Internet. Things. J.* **2026**, *13*, 17003-10. DOI
184. Wang, Y.; Shangguan, Y.; Yang, X.; He, W.; Shi, Y. Resilience analysis of closed-loop multiagent systems under replay attacks in the context of consensus tracking. *IEEE. Trans. Cybern.* **2026**, *56*, 5185-96. DOI
185. Guo, J.; Dou, C.; Yue, D.; Zhang, B.; Zhang, Z.; Zhang, Z. Multi-agent system based cooperative power control of distributed virtual power plants for frequency regulation under diversified cyber attacks. *IEEE. Trans. Power. Syst.* **2026**, *41*, 2118-32. DOI
186. Hu, S.; Liang, J.; Chen, X.; Zhang, J.; Xie, X. Co-design of data-driven attack observer and resilient controller for unknown NCSs under hybrid attacks. *IEEE. Trans. Autom. Control.* **2026**, *71*, 3044-57. DOI

187. Omar, K.; Gómez, J. M. Adversarial machine learning in real-time threat detection a multi-agent defense framework against adaptive cyber attacks. In *2026 2nd International Conference on Computational Intelligence Approaches and Applications (ICCIAA)*, Amman, Jordan. May 12-14, 2026. IEEE; 2026. p. 1-5. DOI

188. Shi, X.; Zhang, C. Pole-dynamics attacks detection in MASs by controller-side identical watermarking. In *2026 5th Conference on Fully Actuated System Theory and Applications (FASTA)*, Qinhuangdao, China. May 22-24, 2026. IEEE; 2026. p. 1-7. DOI

Disclaimer/Publisher's Note: All statements, opinions, and data contained in this publication are solely those of the individual author(s) and contributor(s) and do not necessarily reflect those of OAE and/or the editor(s). OAE and/or the editor(s) disclaim any responsibility for harm to persons or property resulting from the use of any ideas, methods, instructions, or products mentioned in the content.



© The Author(s) 2026. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.