Interdisciplinary Journal of Health, Environment and Computation
ISSN Online: 3108-1444 Vol.: 1, No.: 4 (April – June) 2026. PP. : 30-52

THE EVOLUTION AND ARCHITECTURE OF AGENTIC OPEN SOURCE INTELLIGENCE

Arunanshu Chatterjee[1], Anurag Roy[2]

[1]Department of Electrical Engineering, Jalpaiguri Government Engineering College. West Bengal, India.

[2]Department of Computer Science and Engineering and Artificial Intelligence, Techno India University, West Bengal.

Email: [1]18.arun.c@gmail.com, [2]anuragroy485@gmail.com

https://doi.org/10.65983/ijhec.2026.04.0004

Abstract

This comprehensive review paper examines the ontological shift in Open Source Intelligence (OSINT). For decades, the field was strictly defined by the extraction of actionable intelligence from publicly accessible data, operating on a linear, highly manual pipeline that relied entirely on human cognitive processing. The advent of Agentic Artificial Intelligence (AI) fundamentally displaces this foundational paradigm. By transitioning the discipline from reactive data aggregation to proactive, autonomous execution, agentic architectures redefine the intelligence lifecycle, shifting the cognitive burden of data processing, correlation, and initial synthesis from human operators to autonomous algorithmic systems. To clarify the academic boundaries and methodological intent of this document, it must be explicitly stated that this is a comprehensive review paper rather than a presentation of singular novel empirical research. The objective of this review is to systematically aggregate, synthesize, and critically evaluate the theoretical, mathematical, and architectural state-of-the-art across the rapidly expanding domain of Agentic AI in cybersecurity and OSINT. By analyzing recent academic frameworks published across leading repositories—such as the Institute of Electrical and Electronics Engineers (IEEE), the Association for Computing Machinery (ACM), Springer, and Elsevier—this paper provides a definitive structural analysis of the current landscape. This review specifically addresses the mathematical foundations of autonomous reasoning through Bayesian probabilistic updating, constructs structural models of multi-agent collaborative topologies, evaluates the operational efficacy of specialized cybersecurity frameworks, and thoroughly examines the profound socio-technical friction generated by this paradigm shift, specifically focusing on the barrier of algorithmic trust, legal accountability deficits, and the emerging defensive imperative to track Non-Human Identities (NHIs).

Keywords: Open Source Intelligence (OSINT), human cognitive processing, Agentic Artificial Intelligence, Bayesian probabilistic, Non-Human Identities (NHIs).

Received: February 26, 2026; Revised: May 11, 2026; Accepted : June 05, 2026

1. Introduction to the Ontological Shift in Open Source Intelligence

Historically, the practice of open-source intelligence evolved in direct tandem with the prevailing mediums of global communication. Mid-twentieth-century intelligence methodologies relied heavily on the manual transcription and linguistic analysis of foreign radio broadcasts, print media, and intercepted terrestrial communications. As the digital revolution democratized data publication, OSINT adapted through the deployment of automated scraping tools engineered to ingest massive datasets from surface web search engines, highly dynamic social media firehoses, specialized deep web forums, and the encrypted enclaves of the dark web (Almeida Palmieri et al. 2025; Pastor-Galindo et al. 2020). Despite these dramatic advancements in collection velocity and data aggregation, the analytical phase remained strictly constrained by human cognitive limitations.

The traditional intelligence model suffers from a severe structural bottleneck. Existing software utilities permit analysts to cast extraordinarily wide exploratory nets across the internet, but the subsequent operational requirements—identifying salient digital artifacts, validating heterogeneous source credibility, and synthesizing highly fragmented telemetry into a coherent operational picture—remain heavily manual processes (Almeida Palmieri et al. 2025). Highly trained security professionals routinely expend disproportionate cognitive resources managing complex software interfaces, tuning search parameters, and filtering out statistical noise rather than executing high-level strategic reasoning. Consequently, the traditional sequential stages of planning, collection, processing, analysis, and dissemination are perpetually hindered by the limits of human processing capacity.

Furthermore, traditional OSINT is inherently static. In a standard workflow, a human analyst constructs a query, retrieves the corresponding data, and drafts a comprehensive threat assessment. Because the global information environment is characterized by relentless, asynchronous data generation, these static intelligence assessments begin to degrade in operational relevance the precise moment they are finalized. Prior technological interventions provided only incremental amplification of human capabilities. The introduction of Retrieval-Augmented Generation (RAG) architectures successfully grounded Large Language Models (LLMs) in external factual datasets, significantly reducing algorithmic hallucination by forcing the model to cite retrieved text (Almeida Palmieri et al. 2025; Atlam 2025). Similarly, Narrative Intelligence platforms applied advanced machine learning analytics to detect behavioral patterns and thematic stories within complex media ecosystems. However, these frameworks remained entirely dependent on human initiation and operated orthogonally to established analytical workflows. Operators were still required to manually construct precise search queries, interpret the contextual weight of the retrieved vectors, and mentally map the complex relationships between disparate entities across multiple domains.

Agentic OSINT resolves this structural latency by fundamentally inverting the operational posture. The system no longer waits passively for a human prompt; instead, it autonomously pursues a predefined strategic objective, representing an epistemological shift from software as a tool to software as an active, independent research operator (Almeida Palmieri et al. 2025).

2. Mathematical and Theoretical Foundations of Agentic Reasoning

The transition from static machine learning models and deterministic automation scripts to autonomous agency represents a structural revolution in digital tradecraft (Almeida Palmieri et al. 2025; Abou Ali et al. 2025). At a fundamental level, autonomous intelligence agents are software constructs explicitly engineered to independently interpret high-level semantic objectives, formulate multi-step execution strategies, and actively manipulate external digital environments via Application Programming Interfaces (APIs) and discrete software utilities. Unlike conventional foundational models that map input prompts to output probabilities in a single localized execution, agentic systems possess a complex suite of recursive cognitive loops. To achieve this continuous execution safely and logically, these systems rely on deeply embedded mathematical foundations that map input data to probabilistic states and optimize decision-making trajectories over prolonged operational timelines.

2.1. Bayesian Probabilistic Updating in Intelligence Gathering

The defining technical characteristic of an autonomous OSINT agent is its capacity for continuous, unprompted execution governed by rigorous probabilistic logic. When deployed within persistent cloud environments, agents continuously conduct asynchronous monitoring of highly dynamic data streams, including newly registered domain name registries, certificate transparency logs, and decentralized dark web paste sites (Almeida Palmieri et al. 2025; Sun et al. 2023). To mathematically represent the evolving state of the external environment, agentic OSINT utilizes Bayesian probabilistic updating.

In this mathematical framework, each analytic question or intelligence objective is treated as a hypothesis, denoted as H. For instance, a hypothesis might state, "A specific transnational threat actor is operating this newly discovered command-and-control infrastructure." Before observing new empirical data, an agent's background knowledge and historical intelligence are mathematically codified as the prior probability, $P(H)$ (Palabindela and Konnipati 2026). Upon encountering novel OSINT evidence, denoted as E (such as an intercepted communication, a behavioral anomaly, or a newly scraped IP address), the agent automatically triggers a re-evaluation of its internal intelligence graph. The agent updates its belief utilizing Bayes' theorem:

$$P(H|E)=\frac{P(E|H)P(H)}{P(E)}$$

Within this equation, $P(H|E)$ represents the posterior probability, which is the updated statistical confidence that the hypothesis is true given the new evidence. The term $P(E|H)$ represents the likelihood of observing the specific evidence if the hypothesis were indeed true, while $P(E)$ serves as the marginal probability of the evidence, acting as a normalizing constant to ensure the resulting probabilities sum to one (Palabindela and Konnipati 2026).

Because OSINT investigations involve the continuous ingestion of thousands of disparate, multilingual data points over time, the system applies this mathematical update iteratively. For sequential evidence arriving asynchronously $(E_1, E_2, \dots, E_n)$,

the framework updates the posterior probability dynamically using the following formula:

$$P(H|E_1, \dots, E_n) \propto P(E_n|H)P(H|E_1, \dots, E_{n\text{-}1})$$

This precise formulation allows the intelligence product to cease being a static document; it becomes a continuously evolving probability matrix reflecting the real-time state of the external environment (Almeida Palmieri et al. 2025; Palabindela and Konnipati 2026; Sefati et al. 2025; Sefati 2026).

These probabilistic equations are uniquely adapted to specific investigative tasks. Palabindela and Konnipati detail how these formulas optimize specialized operations (Palabindela and Konnipati 2026). For the task of Entity Disambiguation—resolving an ambiguous digital alias or mention m to a known canonical entity e—the equation is adapted to $P(e|m) \propto P(m|e)P(e)$. Here, the prior $P(e)$ encodes historical analyst expectations or the frequency of occurrences, while the likelihood $P(m|e)$ relies on contextual text similarity cues derived from natural language processing.

Furthermore, for Deception Detection tasks, a Naive Bayes model or a statistical text classifier calculates the probability of a statement being deceptive ($H=1$) or truthful ($H=0$). The agent mathematically flags the content as deceptive if $P(H=1|\text{text})>0.5$. By translating structured analyst traces, such as think-aloud protocols and human intelligence (HUMINT) tip annotations, into parameterized cognitive priors, this Bayesian approach consistently outperforms naive fusion baselines. The quantitative results demonstrated a maximum enhancement of 34 percent in evidence relevance ranking and a 28 percent decrease in misclassification errors across diverse OSINT datasets (Palabindela and Konnipati 2026).

2.2. Decision-Theoretic Control and Expected Utility

Once beliefs are updated via the Bayesian network, the autonomous agent must take action. This action might involve executing a secondary port scan, invoking a reverse-image search API, pivoting to a new social media profile, or escalating a critical alert to a human operator. The transition from internal probabilistic belief to external action is formally governed by Bayesian decision theory (Palabindela and Konnipati 2026).

Given a set of possible actions a and the true state of the world s (e.g., true entity identity, event sequence, or truthful versus deceptive), a utility function $U(a, s)$ mathematically quantifies the inherent value of executing action a if state s holds true. The expected utility of any action under the agent's current belief state is computed as the sum of the utilities of all possible states, weighted by the posterior probability of those states:

$$E[U(a)]=\sum_{s} P\,(s|E)U(a, s)$$

Following the Rational Agent Rule, the autonomous system is programmed to select the optimal action a^* that maximizes expected utility, mathematically expressed as:

$$a^{*}=\underset{a}{argmax}E[U(a)]$$

This decision-theoretic control layer is critical for maximizing expected information gain while simultaneously bounding the agent's actions according to legal, ethical, and computational cost constraints. It permits the encoding of asymmetric cost functions—for instance, mathematically penalizing false alarms differently than missing a true malicious event—ensuring that the agent operates within strictly defined risk tolerances (Palabindela and Konnipati 2026).

2.3. Markov Decision Processes and Autonomous Threat Hunting

While single-step utility maximization is highly effective for localized tasks such as entity extraction or sentiment analysis, complex cybersecurity operations—such as multi-stage threat hunting, dynamic vulnerability exploitation, or lateral network movement—require agents to plan and reason across extended temporal horizons. Consequently, agentic navigation in highly dynamic environments is formally modeled as a Markov Decision Process (MDP).

The continuous learning and decision-making process is governed by a mathematical tuple (S, A, P, R), defined as follows:

- **State Space (S)**: Represents the agent's current intelligence graph, integrating retrieved Cyber Threat Intelligence (CTI) reports, identified Tactics, Techniques, and Procedures (TTPs), network topologies, and available vulnerabilities.
- **Action Space (A)**: Consists of discrete tool invocations available to the agent. This includes executing Python scripts, schema inspection, querying APIs, rotating cloud infrastructure, or interacting with a target system.
- **Transition Probability (P)**: The probability that taking a specific action a while in state s will successfully lead to a new state s'.
- **Reward Function (R)**: The numerical value assigned to discovering high-fidelity intelligence or successfully exploiting a vulnerability, while actively penalizing resource exhaustion, API timeouts, or defensive detection.

The agent mathematically evaluates the optimal path through the environment by computing the Bellman Optimality Equation, which calculates the maximum expected cumulative reward for a given state-action pair:

$$Q^{*}(s, a = R$$

Where γ represents the discount factor, ensuring mathematical convergence over potentially infinite operational horizons by prioritizing immediate intelligence gains over distant, uncertain rewards. In highly obfuscated environments where the agent cannot fully observe the underlying network state—such as mapping dark web infrastructure or navigating encrypted networks—this framework is extended into a Partially Observable Markov Decision Process (POMDP). In a POMDP, the agent maintains a probabilistic belief state over the underlying true state, using incoming observation streams $\omega(t)$ to constantly refine its trajectory toward the optimal action stream $a(t)$.

To establish a baseline for algorithmic comparison, early autonomous defense models relied on deterministic heuristic rules. As formalized in recent research, the *Heuristic Agent Decision Making* algorithm evaluates the state of the network and iteratively applies predefined rulesets to select contextually appropriate defense actions.

Initialize State Space S **Function** HeuristicDecision(Network Telemetry t) $s \leftarrow \text{Parse}(t)$ Identify compromised hosts within s Select contextually appropriate defense action a Execute randomized exploratory scan a **End Function**

While heuristic agents provide a functional baseline, the rigid nature of predefined conditional logic consistently fails to adapt to novel adversarial TTPs, illustrating the absolute necessity of transitioning to continuous, reinforcement-based mathematical optimization through MDPs.

3. Architectural Topologies and Structural Workflows

The empirical efficacy of agentic OSINT relies fundamentally on its underlying architectural topology. Deploying individual agents or massive fleets of specialized worker models requires precise software design patterns that strictly dictate state management, inter-agent communication, and recursive execution loops. To address the structural complexity of these systems, the following architectural representations detail the functional mechanisms driving both single-agent cognition and multi-agent collaboration.

3.1. Single Agent Architecture Model

Operating in isolation, a highly capable autonomous OSINT agent functions via an internal, highly recursive cognitive loop. This architecture abstracts the immense complexity of LLM memory management, context routing, and API tool invocation into a unified execution engine.

Visual Architecture of the Autonomous OSINT Agent.

Structural Representation of the Autonomous OSINT Agent Architecture

Table 1: Architectural Layers and Functional Components of the Proposed AI-Agent Framework

Architectural Layer	Component Designation	Functional Description and Mathematical Role
Perception Layer	Data Ingestion Nodes	Continuously parses unstructured text, multi-modal imagery, and continuous API streams. Interfaces asynchronously with surface web feeds, deep web databases, and encrypted networks.
Cognitive Engine	LLM Reasoning Core	Executes the Bayesian updating calculations and evaluates the Expected Utility $E[U(a)]$ of potential actions. Enforces Chain-of-Thought (CoT) reasoning to mathematically compel step-by-step logical deductions prior to execution.
Memory State	Short & Long-Term Memory	Maintains the dynamically evolving probabilistic hypothesis graph. Tracks context windows and archives historical evidence into vector databases to prevent cyclical algorithmic looping or redundant API calls.
Action & Tool Space	Callable Utilities (A)	Dynamically formulates precise command-line syntax to execute legacy scripts natively, invoking external APIs, and executing deterministic network interactions without requiring manual

Architectural Layer	Component Designation	Functional Description and Mathematical Role
		prompting (Almeida Palmieri et al. 2025).

3.2. Multi-Agent Communication Models

While an individual agent drastically reduces data processing latency, complex intelligence requirements demand highly diverse, specialized skill sets that exceed the capacity of a single generalized model. As analyzed by Abou Ali et al. (Abou Ali et al. 2025), deploying a fleet of narrowly focused, heavily prompted expert agents yields vastly superior analytical precision compared to utilizing a monolithic foundational model. By restricting an agent's operational scope, developers precisely parameterize the optimization functions for specific subtasks.

To operationalize these fleets, engineers rely on distinctly structured communication topologies that govern how agents interact, share state, and divide computational labor.

Multi-Agent Collaborative Topology (Orchestrator-Worker Model).

Structural Representation of Multi-Agent Collaborative Topologies

Table 2: Multi-Agent Collaboration Topologies and Their Operational Characteristics

Topology Model	Collaboration Dynamics	Optimal Operational Use Case
Orchestrator-Worker	Centralized LLM orchestrator delegates to specialized worker agents, synthesizing returned JSON arrays (Almeida Palmieri et al. 2025).	Standard enterprise OSINT pipelines and managing multi-step RAG workflows securely.
Hierarchical	Top-level strategic agent delegates operational domains to manager agents, directing specialized workers.	Tracking state-sponsored espionage across distinct physical, geopolitical, and cyber jurisdictions.
Blackboard	Agents monitor and post asynchronously to a communal knowledge base for non-linear problem solving.	Complex, emergent problem-solving requiring non-linear deductive reasoning, mimicking expert panels.
Market-Based	Subtasks are queued; agents bid mathematically based on computational cost and confidence metrics.	Optimizing large-scale, cost-sensitive distributed cloud infrastructure deployments.

3.3. The Bayesian Update Workflow

To synthesize the mathematical theories and architectural models into functional software, agentic systems execute a rigidly defined, event-driven workflow. This sequential process ensures data provenance, avoiding semantic loss or algorithmic hallucination during complex investigative handoffs.

Continuous Bayesian Update Workflow.

Structural Representation of the Continuous Bayesian Update Workflow

Table 3: Bayesian Decision-Making Workflow of the Proposed AI Agent

Workflow Phase	Action and Trigger	Computational Output
1. Evidence Detection	An anomaly detected across the monitored stream triggers event webhook or Pub/Sub message.	Raw digital evidence (E_n) is ingested into the system.
2. Prior Retrieval	The agent queries the secure vector database for the existing graph.	Prior $P(H)$ is loaded into working memory.
3. Likelihood Computation	Core engine cross-references artifact with threat models and linguistic patterns.	Likelihood $P(E_n \mid H)$ conditional probabilities established.
4. Posterior Integration	The core update equation is executed at machine speed to update confidence weightings (Palabindela and Konnipati 2026).	Evolving posterior matrix $P(H \mid E_1 \dots E_n)$ generated.
5. Action Threshold	System calculates Expected Utility $E[U(a)]$ for action space. If above threshold, agent acts.	Tool invocation, database write, or escalation alert to the operator.

The rapid transition towards multi-agent automation is corroborated by recent advancements across diverse domains. Research on autonomous driving security (Chernikova et al. 2019; Patel et al. 2024; D. Zhao et al. 2025; R. Zhao et al. 2025), federated learning (Ongun et al. 2025), and collaborative edge intelligence (Gupta and Sharma 2026; Jiang 2025; Li et al. 2019; Su et al. 2024; Taleb 2024; Yuan et al. 2024) heavily leverages related mathematical foundations to map agentic control and threat detection. Similarly, complex cyber-physical environments, such as smart manufacturing (Shaik et al. 2023) and maritime IoT (Kavallieratos et al. 2020; Park et al. 2023), utilize hybrid probabilistic networks to quantify adversarial dynamics. Furthermore, studies focusing on the structural resilience of cloud services (Caviglione et al. 2021; Tunc et al. 2019), dataset poisoning mitigations (Jagielski et al. 2021; Severi et al. 2025; Suri et al. 2026; Venkatesan et al. 2021), and foundational LLM privacy vulnerabilities (Chaudhari et al. 2025, 2026; Naseh et al. 2025) provide the necessary empirical grounding to ensure agentic operations remain securely bounded. These are reinforced by innovations in drone network defenses (Tunc 2025), zero trust engines (Tunc 2023), reconfigurable intelligent surfaces (Liu et al. 2024), 3D passenger tracking (Liu et al. 2023), system-of-systems SOA frameworks (Mohsin and Janjua 2018; Mohsin et al. 2019; Teixeira et al. 2020), blockchain risk modeling (Feng et al. 2021), risk evaluations (Avalos and Tunc 2025; Ouyang et al. 2026), secure ERP release pipelines (Gangaiah et al. 2026), early DevSecOps testing (Thanvi et al. 2026), and cloud analytics breach detection (Tiwari et al. 2019).

By utilizing distributed cloud infrastructure such as serverless compute environments and high-throughput data streaming pipelines, enterprises successfully orchestrate these asynchronous interactions. This infrastructure provides vital fault tolerance, ensuring that localized failures—such as a worker agent encountering an unexpected

API timeout or a CAPTCHA block—do not collapse the broader multi-step investigation.

4. Academic Innovations and Formal OSINT Frameworks

The theoretical potential of integrating multi-agent reasoning, continuous Bayesian updating, and dynamic tool orchestration has been formally codified in recent, highly rigorous academic architectures. The rapid transition from experimental sandbox environments to mission-grade deployment is characterized by systems explicitly engineered to address the distinct forensic and operational requirements of intelligence operations.

4.1. The Unified Generative and Agentic OSINT Architecture

In 2025, researchers proposed a comprehensive, unified agentic architecture explicitly designed to overcome the critical scalability, adaptability, and forensic limitations inherent in traditional script-based intelligence pipelines (Almeida Palmieri et al. 2025).

The framework conceptualizes the OSINT lifecycle not as a sequence of manual tasks, but as a fully integrated, modular ecosystem. Crucially, the architecture mandates the deep integration of Retrieval-Augmented Generation (RAG) to enforce strict factual grounding, tightly paired with Chain-of-Thought (CoT) reasoning protocols. By mathematically compelling the underlying LLM to articulate its logical deductions step-by-step prior to executing any external action, the framework drastically reduces algorithmic hallucination while establishing a highly transparent forensic audit trail of its analytical outputs.

The framework is structured across four highly specialized, interacting operational modules:

1. **Multi-source Data Ingestion**: Responsible for the continuous, asynchronous collection of unstructured, multimodal data across diverse surface web streams, deep web databases, and encrypted dark web networks, ensuring a persistent feed of localized global telemetry (Almeida Palmieri et al. 2025).
2. **LLM-Powered Analysis**: Serving as the primary cognitive engine, this module leverages advanced natural language processing to translate cross-lingual intercepts, extract complex entity relationships, and mathematically interpret the contextual sentiment and adversarial intent of raw intelligence.
3. **Generative Scenario Simulation**: A uniquely advanced capability that utilizes generative AI to mathematically model and simulate adversarial scenarios. This allows the system to synthetically generate potential threat actor reactions to various defensive postures, effectively pressure-testing competing investigative hypotheses before committing physical or digital resources.
4. **Ethical Safeguard Enforcement**: Embeds structural transparency logging mechanisms, mandatory bias detection algorithms, and critical human-in-the-loop validation checkpoints, ensuring that autonomous operations remain legally compliant and strictly proportional.

To interface effectively with the external environment, the architecture utilizes a modular orchestration layer featuring comprehensive tool integration, granting agents

programmatic access to infrastructure mapping utilities like Shodan, real-time social media APIs, and dark web Threat Intelligence feeds. The operational viability of this unified architecture was empirically validated through a complex proof-of-concept focused on locating missing persons using strictly publicly available data. The autonomous operation initiated continuous environmental monitoring, correlated faint multimodal digital artifacts, executed rigorous credibility assessments to filter out intentional disinformation, and generated probabilistic location models. The results demonstrated massive, measurable improvements in intelligence coverage, analytical accuracy, and decision speed when compared directly to conventional, human-led workflows.

Open Source Frameworks and Optimal Use Cases

Table 4: Open-Source Agentic AI Frameworks and Their Architectural Characteristics

Open Source Framework	Architectural Focus and Optimal Use Cases
Dify	Focuses on comprehensive LLM gateways and complex visual workflow orchestration. Widely utilized for deploying autonomous financial analysis suites and managing multi-step RAG pipelines.
AutoGen	Developed by Microsoft Research, this framework excels in orchestrating multi-agent conversational ecosystems. It enables agents to challenge and refine each other's outputs through structured dialogue, replacing monolithic control with emergent problem-solving.
CrewAI	A specialized role-playing framework enabling developers to strictly define agent personas (e.g., senior researcher, forensic analyst, technical writer) that collaborate sequentially or hierarchically to execute operations.
LangGraph	An extension of the LangChain ecosystem that treats agent workflows as highly cyclical Directed Acyclic Graphs (DAGs). Optimal for applications demanding rigorous state management and persistent memory loops.
OpenAI Agents SDK	A lightweight Python framework designed for rapid prototyping, seamless API integration, and the deployment of general-purpose autonomous agents.

4.2. Specialized Cybersecurity Frameworks: The CAI Architecture

While OSINT inherently focuses on passive data aggregation and analysis, executing offensive or defensive maneuvers within secure network environments demands highly specialized programmatic capabilities. General-purpose orchestrators manage cognitive routing, but specialized frameworks grant agents the exact utilities required to manipulate secure endpoints. The Cybersecurity AI (CAI) framework represents a paradigm-defining advancement in this highly technical domain.

Developed as a modular, lightweight, open-source system explicitly designed to empower security professionals, the CAI architecture bridges the gap between raw LLM reasoning and practical, bug-bounty-ready security testing (Mayoral-Vilches et al. 2025). The framework operates on an agent-centric design that supports over 300 distinct foundation models, equipping specialized agents with a massive arsenal of built-in cybersecurity tools. CAI enables the rapid instantiation of autonomous entities dedicated to discrete tasks such as initial network reconnaissance, dynamic vulnerability discovery, active exploitation, and privilege escalation. To mitigate the

profound risks associated with autonomous execution in secure environments, CAI integrates rigorous internal guardrails designed to detect adversarial prompt injection and rigidly prevent the execution of catastrophic, system-breaking commands (Almeida Palmieri et al. 2025; Mayoral-Vilches et al. 2025).

The empirical efficacy of the CAI framework has been proven in highly rigorous international cyber-competitions. CAI-powered autonomous agents achieved a Top-10 global ranking in the competitive Dragos Operational Technology (OT) Capture The Flag (CTF) competition, autonomously completing 32 of 34 complex challenges and maintaining a 37 percent accuracy rate across diverse industrial control system protocols (Almeida Palmieri et al. 2025; Mayoral-Vilches et al. 2025). By enabling non-professionals to discover significant security bugs (CVSS 4.3–7.5) at rates comparable to experts, the framework serves as a democratizing force, challenging the oligopolistic ecosystem historically dominated by major bug bounty platforms.

Furthermore, the operational reach of agentic systems has demonstrably extended beyond theoretical network testing into high-impact cyber-physical interaction. In a subsequent study, researchers documented a systematic security assessment of three diverse consumer robotic platforms: an autonomous lawnmower, a powered exoskeleton, and a window-cleaning robot (Mayoral-Vilches et al. 2026). Operating entirely autonomously, the CAI agents uncovered 38 distinct vulnerabilities (including 16 critical flaws) across the platforms, exploiting fundamental authentication failures such as unauthenticated Android Debug Bridge (ADB) instances, unprotected Bluetooth Low Energy (BLE) commands, and default administrative credentials. These empirical findings unequivocally establish that AI agents now possess the mechanical intelligence required to autonomously replicate and extend vulnerability discovery in complex physical systems, shifting the balance of power asymmetrically toward automated offensive capabilities and necessitating a rapid evolution toward GenAI-native defensive architectures like the Robot Immune System (RIS) (Mayoral-Vilches et al. 2026).

4.3. Programmatic Assimilation of Legacy OSINT Utilities

The advent of agentic reasoning does not render the vast, established ecosystem of legacy OSINT tools obsolete; rather, it fundamentally recontextualizes how they are operated. Traditional command-line Python scripts, historically designed for discrete, manual data gathering, are seamlessly assimilated into the agent's architecture as callable "skills" or "tools" constituting the action space A.

Legacy OSINT Tools and Agentic Functionality

Table 5: Legacy OSINT Tools and Their Agentic Functional Roles

Legacy OSINT Tool	Agentic Functionality within Workflow
holehe	Agents autonomously invoke this script to query hundreds of global platforms to determine if a specific target email address is registered, exploiting forgotten password metadata functionalities.
Toutatis	Invoked autonomously by agents to extract deeply obfuscated metadata from Instagram accounts, pulling hidden email addresses and phone numbers.
Spyder OSINT	Utilized autonomously for sequential phone number reverse lookups, IP address geolocation, and cross-platform username verification.

Legacy OSINT Tool	Agentic Functionality within Workflow
Mr. Holmes / Seekr	Multi-purpose information gathering toolkits that agents interface with to execute comprehensive routines spanning multiple social networks.

For example, when an autonomous researcher agent isolates a target email address during an investigation, it does not pause the execution loop to alert a human operator. Leveraging its tool-calling capabilities, the agent dynamically formulates the correct command-line syntax and autonomously executes legacy scripts—such as *holehe*, which exploits forgotten password metadata to verify email registration across hundreds of global platforms, or *Toutatis*, which extracts deeply obfuscated metadata from Instagram accounts. The agent mathematically parses the resulting JSON standard output within a sandboxed environment, extracts confirmed social media profiles, and appends this new intelligence to its dynamic probability graph before mathematically reasoning about the next logical step. By chaining these utilities autonomously—piping the output of an initial discovery directly into the parameters of a secondary exploitation tool—agents execute complex, multi-step exploitation chains end-to-end with unprecedented speed and precision.

5. Tiered Autonomy and Security Operations Center Integration

As agentic intelligence capabilities scale dynamically, the operational deployment of these autonomous systems within high-stakes environments—specifically Security Operations Centers (SOCs)—must be strictly governed by progressive tiers of autonomy. Unchecked autonomous execution presents unacceptable systemic risks to critical infrastructure. Therefore, the architectural transition from assistive software tools to autonomous collaborative partners requires rigorous governance frameworks that map directly to established operational risk parameters.

5.1. Progressive Autonomy and Escalation Tiers

To maintain systemic stability, AI agents are integrated into SOC environments via three distinct operational tiers:

- **Tier 1 (Augmented Triage)**: At this foundational level, agentic systems operate with strictly limited autonomy, serving primarily as advanced, high-speed triage engines. These agents execute the automated classification, mathematical deduplication, and contextual enrichment of incoming security telemetry. By continuously correlating raw network alerts against historical internal incident data and global OSINT threat feeds using Bayesian updating, Tier 1 agents surface high-fidelity context, insulating human analysts from pervasive alert fatigue. Crucially, while the AI calculates the probability of malicious intent and mathematically recommends containment actions—such as host isolation or account lockdown—it requires explicit, manual human authorization before executing any active network intervention.
- **Tier 2 (Deterministic Execution)**: Tier 2 autonomy introduces the capability for deterministic execution within strictly constrained, pre-authorized environments. Operating under the boundaries of rigid incident response playbooks, these agents are seamlessly integrated with Security Orchestration, Automation, and Response (SOAR) frameworks. Once a specific mathematical threshold of malicious activity is definitively verified by the cognitive engine, a Tier 2 agent autonomously executes predefined actions—

such as isolating a compromised endpoint from the corporate network or immediately revoking compromised cryptographic certificates. This capability massively accelerates incident response times while maintaining a firm boundary around critical infrastructure.

- **Tier 3 (Exploratory Threat Hunting)**: Tier 3 represents the current vanguard of deployed agentic capabilities, engineered to support highly advanced, exploratory threat analysis. At this operational level, sophisticated systems autonomously execute deep threat hunting, dynamic vulnerability scanning, and multi-stage root cause analysis. These agents proactively navigate complex network topologies, correlate unstructured telemetry across heterogeneous global systems, and generate comprehensive summary reports containing proposed structural mitigation architectures.

5.2. The Human-in-the-Loop Collaborative Framework

The deployment of Tier 3 autonomy necessitates a fundamental evolution in sociotechnical architectures. Because advanced cyber-attacks evolve rapidly, human oversight remains critical to differentiate highly privileged, legitimate administrative actions from deeply obfuscated, sophisticated malicious behavior. Contemporary deployments tether orchestrator agents to stringent Human-in-the-Loop (HITL) validation mechanisms, ensuring the system operates within statistically validated safety margins.

Human-in-the-Loop (HITL) Collaborative SOC Framework.

Structural Representation of the HITL Collaborative SOC Framework

Table 6: Human–AI Interaction Framework for AI-Driven Cyber Threat Intelligence

Framework Pillar	Functional Mechanism and Human-AI Interaction Dynamics
API-Driven Orchestration	Autonomous agents seamlessly integrate with SIEMs, firewalls, and EDR platforms, gathering multi-source data to build the initial intelligence state (Mohsin et al. 2025).
Situational Awareness	The AI engine executes rapid incident enrichment and correlates heterogeneous telemetry, translating massive datasets into coherent visual assessments for the human operator.
Bidirectional Collaboration	Continuous feedback mechanism where human analysts evaluate the AI's proposed "plan correction" features. The agent must formulate and preview a detailed execution plan before altering network states (Mohsin et al. 2025).
Trust and Accountability	Human operators serve as the ultimate decision nodes for novel threat vectors, maintaining legal accountability and steering the overarching strategic investigation (Mohsin et al. 2025).

At its core, this framework establishes that a bidirectional collaboration loop between AI scale and human intuition—mediated by advanced situational awareness and strict algorithmic trust—is the only viable mechanism to enable adaptive decision-making and feedback-driven learning in modern cybersecurity operations (Mohsin et al. 2025).

5.3. Explainable AI as a Mechanism for Operational Trust

The foundational barrier to integrating autonomous agents into SOCs is the paradigm of algorithmic trust. In high-stakes environments, analysts cannot rely solely on empirical benchmarks of general AI accuracy; they must possess absolute confidence that the agent's emergent reasoning aligns with their operational intent and tactical judgment. Trust is deeply dependent on transparency. Models that function as impenetrable "black boxes" inevitably foster suspicion and operational rejection.

Explainable AI (XAI) seeks to resolve this friction by mathematically exposing the internal logic of complex neural architectures. Methodologies such as SHAP (SHapley Additive exPlanations) provide vital model-agnostic interpretability, allowing operators to visually inspect the exact statistical weight the algorithm assigned to specific variables during the decision-making process. By demystifying the algorithmic reasoning process, XAI reduces cognitive overload in Security Operations Centers, enabling analysts to rapidly validate conclusions and allocate investigative resources with maximum efficiency. To foster genuine trust, agentic platforms must supply end-to-end traceability, logging every API invocation, tool usage, and hypothesis generation, mathematically mapping them back to the source material to ensure irrefutable data provenance.

6. Managing Non-Human Identities and Cryptographic Governance

The operational autonomy of agentic OSINT and cyber-defense systems is inextricably linked to their capability to interact dynamically with global digital infrastructure. To navigate complex networks, dynamically provision and rotate anonymous cloud infrastructure, and aggressively manage persistent access vectors, autonomous agents cannot rely on human operators to manually input credentials. Consequently, agents are provisioned with their own cryptographic service accounts, OAuth tokens, API keys, and distinct login credentials (Almeida Palmieri et al. 2025). They function across the internet as highly active, independent Non-Human Identities (NHIs).

6.1. The Identity Sprawl and Zero-Trust Workload Architectures

The aggressive proliferation of these autonomous entities introduces massive identity sprawl within enterprise ecosystems. Traditional Identity and Access Management (IAM) systems were structurally engineered for static human behavioral patterns and implicitly trusted perimeter models; they are fundamentally incapable of governing autonomous, ephemeral workloads that evaluate and execute decisions at machine speed. If an adversary compromises the static API key of a highly privileged intelligence agent, they inherit the entirety of its action space A, blending seamlessly into normal machine-to-machine (M2M) traffic and bypassing traditional network defenses (Almeida Palmieri et al. 2025; Kupris et al. 2025; Banerjee and Singh 2025).

To mitigate this catastrophic systemic vulnerability, modern enterprise architectures are rapidly transitioning from static credential issuance to dynamic, workload-identity models grounded strictly in Zero Trust Architectures (ZTA). The foundational principle of ZTA—"never trust,

always verify"—is highly effective when applied to non-human entities via micro-segmentation and continuous cryptographic validation.

Reliance on API keys must be eliminated to secure multi-agent ecosystems (Pappu et al. 2025). Engineers architect solutions leveraging the Secure Production Identity Framework for Everyone (SPIFFE). In this advanced paradigm, agents do not possess static, long-lived secrets. Instead, a central SPIRE Server acts as the certificate authority, issuing short-lived X.509 SVID certificates—configured, for instance, with a one-hour Time-To-Live (TTL)—based on continuous workload attestation (Pappu et al. 2025).

All inter-agent communication is rigorously authenticated through mutual TLS (mTLS). During the TLS handshake, each agent mathematically verifies the peer's certificate chain against the SPIRE trust bundle and extracts the specific SPIFFE ID from the Subject Alternative Name (SAN) extension to enforce granular, identity-based authorization. This automated cryptographic rotation eliminates the operational burden of manual key management while providing robust security guarantees that mitigate workload impersonation, man-in-the-middle attacks, and unauthorized LLM access (Pappu et al. 2025).

6.2. Identity Threat Detection and Response

From a defensive intelligence perspective, the weaponization of NHIs by threat actors necessitates a complete, structural evolution in digital threat attribution. Traditional OSINT identity resolution relied heavily on tracking the operational security failures of human targets—a reused password across dark web forums, an accidental IP address leak, or a linguistic anomaly. As adversaries increasingly deploy agentic AI to conduct their own automated reconnaissance, this human-centric attribution model collapses. Defensive OSINT must adapt to track entirely algorithmic behaviors.

This critical requirement has catalyzed the emergence of Identity Threat Detection and Response (ITDR) as a formalized security discipline. ITDR is engineered to detect and mitigate identity-based threats before escalation (Kupris et al. 2025). By conducting a rigorous analysis of the MITRE ATT&CK framework, researchers successfully identified 366 specific identity-related threats, developing a structured mechanism to categorize their impact across the digital identity lifecycle.

Modern AI-powered ITDR solutions leverage unsupervised machine learning, sequential modeling of session dynamics, and temporal graph networks to continuously monitor the behavior of non-human identities. By establishing mathematical baselines for an agent's standard API tool-chain preferences, typical infrastructure rotation speeds, and volumetric execution patterns, ITDR systems can instantly detect anomalous programmatic access requests (Almeida Palmieri et al. 2025; Kupris et al. 2025). When an identity threat is detected, ITDR seamlessly integrates with SOAR platforms to initiate automated incident responses—such as real-time access revocation or

immediate session termination—drastically reducing the dwell time of compromised agents and closing the visibility gaps left by traditional Endpoint Detection and Response (EDR) platforms (Kupris et al. 2025).

Furthermore, extending identity trust into highly vulnerable, resource-constrained environments—such as the Industrial Internet of Things (IIoT)—requires decentralized innovations. Research demonstrated that combining public and private blockchain layers ensures both operational privacy and regulatory transparency, effectively mitigating the risks of compromised NHIs altering critical physical infrastructure (Banerjee and Singh 2025).

7. Governance: Algorithmic Trust, Ethics, and Legal Accountability

The undeniable operational efficiency of agentic systems—processing vast datasets, formulating multi-step exploitation chains, and updating probability matrices at machine speed—generates profound socio-technical and legal friction. As intelligence agencies and corporate enterprises aggressively integrate these autonomous entities into their core workflows, the rapid deployment drastically outpaces the structural capacity of existing legal, ethical, and governance frameworks.

7.1. The Cultural and Statistical Dimensions of Algorithmic Trust

The aggressive integration of highly autonomous systems into high-stakes environments introduces the fundamental barrier of algorithmic trust. In arenas encompassing national security, military force protection, and global financial intelligence, operators simply cannot rely on generalized empirical benchmarks (Almeida Palmieri et al. 2025). They must possess unshakeable confidence that the agentic system inherently acts in accordance with strict legal standards and ethical proportionality.

Trust, however, is not a universally standardized metric; it is highly dependent on statistical literacy and cultural context. Recent empirical analyses examining the cross-cultural framework for algorithmic trust reveal that user confidence is intrinsically tied to data transparency mechanisms. Findings demonstrate that cultural adaptation is crucial; collectivistic cultures exhibit markedly different preferences for social validation within algorithmic explanations compared to individualistic markets. Furthermore, statistical literacy plays a complex role. Studies indicate that high statistical literacy is negatively associated with blind trust in algorithms during high-stakes situations, as literate individuals utilize their knowledge to critically evaluate the algorithmic outputs against external realities rather than passively accepting the machine's conclusions. To build sustainable algorithmic trust, agentic platforms must prioritize comprehensive governance, ensuring models are devoid of training bias and capable of contextually adapting their transparency mechanisms to the specific needs of diverse global stakeholders.

7.2. The Accountability Deficit and International Law

The precise, autonomous efficiency of these agents creates a profound and highly dangerous accountability deficit (Almeida Palmieri et al. 2025). When an autonomous AI agent independently flags a civilian for enhanced digital scrutiny, mathematically recommends a surveillance operation, or drafts an intelligence assessment that directly shapes ministerial decision-making, the foundational premises of proportionality and fundamental human rights protection begin to collapse. The structural danger lies in the inherent inversion of accountability: the machine executes the high-level cognitive reasoning, and the human operator merely provides a procedural sign-off, subtly smuggling highly complex, automated biases into sensitive spaces where legal proportionality depends entirely on deliberate human evaluation.

This accountability crisis is heavily scrutinized in adjacent fields experiencing rapid automation, specifically in the deployment of Autonomous Weapon Systems (AWS) in kinetic warfare. The incapacity of current legal frameworks to attribute moral responsibility when autonomous systems execute actions without meaningful human control is a critical issue (Ahmad et al. 2025). AWS fundamentally undermines moral accountability in war, exacerbating risks to civilians and corroding human agency in lethal decision-making, demonstrating that autonomous systems inherently lack ethical legitimacy within the framework of just warfare (Ahmad et al. 2025).

Similarly, analyses highlight the systemic trade-offs between computational speed and contextual discernment (Dibsdale 2025). Research demonstrates that autonomous algorithms inevitably sacrifice nuanced contextual interpretation in highly ambiguous scenarios—precisely where International Humanitarian Law (IHL) demands careful human judgment to uphold the principles of distinction and precaution (Almeida Palmieri et al. 2025; Dibsdale 2025). The deployment of autonomous agents allows operators to emotionally disconnect from the consequences of their intelligence directives, raising profound ethical concerns regarding the total removal of accountability for distant, automated operations (Dibsdale 2025).

Traditional legal concepts concerning liability are structurally inadequate for regulating AI. Concepts such as *mens rea* (criminal intent) and *actus reus* (the act itself) fail to map cleanly to the probabilistic, mathematical outputs of autonomous neural networks (Almeida Palmieri et al. 2025; Mayoral-Vilches et al. 2025). The semi-autonomous nature of these systems severely complicates the identification of responsible parties when agents produce harmful outcomes without direct human intervention. To ensure justice, legal certainty, and continued innovation, legal scholars argue that regulatory systems must evolve beyond traditional agent-centric models. Proposed legal strategies include the adoption of strict liability frameworks, war torts, vicarious liability, and the highly debated concept of establishing a *sui generis* legal personality for sophisticated AI systems to address the immediate

accountability gaps (Ahmad et al. 2025). Without binding international treaties and robust, transparent governance mechanisms that enforce meaningful human control over agentic workflows, the rapid deployment of autonomous OSINT and cyber-agents risks severe non-compliance with fundamental human rights law and exposes institutions to unmanageable legal liabilities (Ahmad et al. 2025).

8. Conclusion

The evolution of Open Source Intelligence from traditional, manual data aggregation into a highly automated, proactive, and agentic discipline represents a definitive paradigm shift in the mechanics of global digital analysis. As this comprehensive review has demonstrated, the deep integration of mathematical models—specifically continuous Bayesian probabilistic updating and Markov Decision Processes—alongside sophisticated, multi-agent collaborative topologies has successfully decoupled intelligence gathering and complex problem-solving from the rigid constraints of human cognitive limits. Academic innovations and open-source frameworks, most notably the Unified Generative OSINT Architecture and the bug-bounty-ready Cybersecurity AI (CAI) framework, have empirically proven that autonomous systems can successfully navigate complex digital networks and exploit vulnerabilities within cyber-physical systems, consistently outperforming traditional human-led methodologies in both speed and operational precision.

However, this transition introduces unparalleled socio-technical complexities that extend far beyond raw computational power. The weaponization of Non-Human Identities (NHIs) necessitates a rapid industry-wide shift toward Zero-Trust workload authentication protocols, such as SPIFFE/SPIRE, and the aggressive adoption of sophisticated Identity Threat Detection and Response (ITDR) platforms to combat entirely algorithmic adversaries. Most critically, the deployment of agentic OSINT confronts deep systemic barriers regarding algorithmic trust, revealing a profound and dangerous legal accountability deficit that current international laws are ill-equipped to handle. As the capabilities of autonomous systems continue to scale dynamically, the primary focus of the academic, cybersecurity, and legal communities must transition from merely expanding operational autonomy toward establishing rigorous, mathematically verifiable frameworks for Explainable AI (XAI) and ethical governance. Only by structurally aligning the operational mathematics of autonomous agents with robust, human-in-the-loop oversight frameworks can the transformative strategic potential of Agentic OSINT be fully realized without compromising the foundational integrity, proportionality, and accountability of global intelligence operations.

Conflict of interest: *Authors are declared that there is no conflict of interest regarding this study.*

References

1. Abou Ali, Mohamad, Fadi Dornaika, and Jinan Charafeddine. 2025. "Agentic AI: A Comprehensive Survey of Architectures, Applications, and Future Directions." *Artificial Intelligence Review* 59 (1): 11. 10.1007/s10462-025-11422-4.
2. Ahmad, Ibrar, Laila Ahmad, Naila Irshad, and Muhammad Talha. 2025. "Artificial Intelligence in Autonomous Weapon Systems: Legal Accountability and Ethical Challenges." *Journal of Engineering, Science and Technological Trends* 2 (1): 120–35.
3. Almeida Palmieri, Eduardo, Mohamed Chahine Ghanem, Viktor Sowinski-Mydlarz, and Dipo Dunsin. 2025. "A Framework for Embedding Generative and Agentic AI in Open Source Intelligence." *Proceedings of the 2025 7th International Conference on Blockchain Computing and Applications (BCCA)*, 838–44. 10.1109/bcca66705.2025.11229637.
4. Atlam, Hany F. 2025. "LLMs in Cyber Security: Bridging Practice and Education." *Big Data and Cognitive Computing* 9 (7): 184. https://doi.org/10.3390/bdcc9070184.
5. Avalos, V. M., and C. Tunc. 2025. "Trust Evaluation Based on a Risk Assessment and Mitigation Framework." *Proceedings of the IEEE International Conference on Smart Internet of Things (SmartIoT)*, 145–52.
7. Banerjee, Tuhin, and Harpreet Singh. 2025. "Securing Non-Human Identities in Industrial IoT: A Blockchain-Based Trust Framework." *NIPES Journal of Science and Technology Research* 7 (4): 228–46. 10.37933/nipes/7.4.2025.1660.
8. Caviglione, L. et al. 2021. "Deep Reinforcement Learning for Multi-Objective Placement of Virtual Machines in Cloud Datacenters." *Soft Computing* 25: 12569–88. 10.1007/s00500-021-06041-3.
9. Chaudhari, Harsh, Jamie Hayes, Matthew Jagielski, Ilia Shumailov, Milad Nasr, and Alina Oprea. 2025. "Cascading Adversarial Bias from Injection to Distillation in Language Models." *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*.
10. Chaudhari, Harsh, Giorgio Severi, John Abascal, et al. 2026. "Phantom: General Trigger Attacks on Retrieval Augmented Language Generation." *ACM Transactions on AI Security and Privacy*.
11. Chernikova, Alesia, Alina Oprea, Cristina Nita-Rotaru, and BaekGyu Kim. 2019. "Are Self-Driving Cars Secure? Evasion Attacks Against Deep Neural Networks for Self-Driving Cars." *IEEE Workshop on the Internet of Safe Things*.

12. Dibsdale, Lucy. 2025. "Beyond Human Judgment – the Morality of Machines: A Critical Examination of the Ethical and Moral Dimensions of Autonomous and AI Weapon Systems in Modern Warfare." *Journal of Global Faultlines* 12 (2): 129–49. 10.13169/jglobfaul.12.2.0002.

13. Feng, Shaohan, Wenbo Wang, Zehui Xiong, Dusit Niyato, Ping Wang, and Shaun Wang. 2021. "On Cyber Risk Management of Blockchain Networks: A Game Theoretic Approach." *IEEE Transactions on Services Computing* 14 (5): 1492–504. 10.1109/TSC.2021.3054223.

14. Gangaiah, Y. K., K. Pappu, and Y. S. Thanvi. 2026. "DevSecOps-Driven Security Controls for ERP Release Pipelines." *Proceedings of the IEEE International Symposium on Digital Forensics and Security (ISDFS)*, 1–6.

15. Gupta, A., and R. Sharma. 2026. "Knowledge Distillation Techniques for Deploying LLMs on Resource-Constrained Vehicular Nodes." *ACM Transactions on Intelligent Systems*.

16. Jagielski, Matthew, Giorgio Severi, Niklas Pousette-Harger, and Alina Oprea. 2021. "Subpopulation Data Poisoning Attacks." *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*.

17. Jiang, Y. 2025. "Edgeshard: Efficient LLM Inference via Collaborative Edge Computing." *IEEE Internet of Things Journal* 12 (10): 13119–31. 10.1109/JIOT.2025.1311931.

18. Kavallieratos, Georgios, Vasiliki Diamantopoulou, and Sokratis K. Katsikas. 2020. "Shipping 4.0: Security Requirements for the Cyber-Enabled Ship." *IEEE Transactions on Industrial Informatics* 16 (8): 11030–40. 10.1109/TII.2020.3012345.

19. Kupris, Erwin, Vitali Serzantov, and Thomas Schreck. 2025. "Identity Threats and Where to Find Them: Mapping ITDR and MITRE ATT&CK." *Proceedings of the 2025 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, 485–92. 10.1109/EDCC66201.2025.00024.

20. Li, Y., S. Ren, P. Wu, S. Chen, C. Feng, and W. Zhang. 2019. "Learning Distilled Collaboration Graph for Multi-Agent Perception." *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

21. Liu, Q., J. Mu, D. Chen, R. Zhang, Y. Liu, and T. Hong. 2024. "LLM Enhanced Reconfigurable Intelligent Surface for Energy-Efficient and Reliable 6G IoV." *IEEE Transactions on Vehicular Technology*.

22. Liu, Y., Q. Lyu, J. Zhang, et al. 2023. "APC: Automatic Passenger Counting and Seat Occupancy Detection for Sightseeing Tram Based on 3D LiDAR and Camera." *Proceedings of the IEEE International Conference on Cybernetics and Intelligent Systems*.

23. Mayoral-Vilches, Víctor, Unai Ayucar Carbajo, O. Laflamme, et al. 2026. "Cybersecurity AI: Hacking Consumer Robots in the AI Era." *arXiv Preprint arXiv:2603.08665*, ahead of print. 10.48550/arXiv.2603.08665.

24. Mayoral-Vilches, Víctor, Luis Javier Navarrete-Lozano, María Sanz-Gómez, et al. 2025. "CAI: An Open, Bug Bounty-Ready Cybersecurity AI." *arXiv Preprint arXiv:2504.06017*, ahead of print. 10.48550/arXiv.2504.06017.

25. Mohsin, Ahmad, Helge Janicke, Ahmed Ibrahim, Iqbal H. Sarker, and Seyit Camtepe. 2025. "A Unified Framework for Human AI Collaboration in Security Operations Centers with Trusted Autonomy." *arXiv Preprint arXiv:2505.23397*, ahead of print. 10.48550/arXiv.2505.23397.

26. Mohsin, A., and N. K. Janjua. 2018. "A Review and Future Directions of SOA-Based Software Architecture Modeling Approaches for System of Systems." *Service Oriented Computing and Applications* 12: 40–56. 10.1007/s11761-018-0240-9.

27. Mohsin, A., N. K. Janjua, S. M. S. Islam, and V. V. G. Neto. 2019. "Modeling Approaches for System-of-Systems Dynamic Architecture: Overview, Taxonomy and Future Prospects." *Proceedings of the 14th IEEE International Conference on System of Systems Engineering (SoSE)*, 24–32.

28. Naseh, Ali, Yuefeng Peng, Anshuman Suri, Harsh Chaudhari, Alina Oprea, and Amir Houmansadr. 2025. "Riddle Me This! Stealthy Membership Inference for Retrieval-Augmented Generation." *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*.

29. Ongun, Talha, Simona Boboila, Alina Oprea, Tina Eliassi-Rad, Jason Hiser, and Jack W. Davidson. 2025. "CELEST: Federated Learning for Globally Coordinated Threat Detection." *IEEE Transactions on Information Forensics and Security*.

30. Ouyang, M. et al. 2026. "Explainable AI for Real-Time Fault Detection." *IEEE Transactions on Reliability*.

31. Palabindela, Sairam, and Sai Madhuri Konnipati. 2026. "Beyond the Data: Bayesian Cognitive Priors for Human-Centered OSINT Automation." *Proceedings of the First International Conference on Advances in Forensics and Cyber Technologies (ICFACT 2025)*, 373–79. 10.2991/978-94-6239-610-4_32.

32. Pappu, Karthik, Badal Bhushan, and Akshay Mittal. 2025. "SPIFFE-Based Zero-Trust Authentication for AI Agent Ecosystems." *Proceedings of the 2025 International Conference on Computer and Applications (ICCA)*, 1–7.

33. Park, Chagnki, Christos A. Kontovas, Chia-Hsun Chang, and Zaili Yang. 2023. "A BN Driven FMEA Approach to Assess Maritime Cybersecurity Risks." *Ocean & Coastal Management* 235: 106511. 10.1016/j.ocecoaman.2023.106511.

34. Pastor-Galindo, J., P. Nespoli, F. G. Marmol, and G. M. Perez. 2020. "The Not yet Exploited Goldmine of OSINT: Opportunities, Open Challenges and Future Trends." *IEEE Access* 8: 10282–304. 10.1109/ACCESS.2020.2965257.

35. Patel, N. et al. 2024. "Privacy-Preserving Federated Learning in Autonomous Vehicle Fleets." *IEEE Transactions on Intelligent Transportation Systems*.

36. Sefati, Seyed Salar. 2026. "Adaptive QoS-Aware Service Composition in the Internet of Things Using a Hybrid Bayesian Network-Based Optimization Algorithm." *IEEE Internet of Things Journal*, ahead of print. 10.1109/JIOT.2026.3678869.

37. Sefati, Seyed Salar, Bahman Arasteh, and Simona Halunga. 2025. "A Comprehensive Survey of Cybersecurity Techniques Based on Quality of Service (QoS) on the Internet of Things (IoT)." *Cluster Computing*, ahead of print. 10.1007/s10586-025-04535-x.

38. Severi, Giorgio, Simona Boboila, John Holodnak, Kendra Kratkiewicz, Rauf Izmailov, and Alina Oprea. 2025. "Model-Agnostic Clean-Label Backdoor Mitigation in Cybersecurity Environments." *Proceedings of the IEEE Military Communications Conference (MILCOM)*.

39. Shaik, S., C. Tunc, and K. Morozov. 2023. "Intrusion Detection for Additive Manufacturing Systems and Networks." *Proceedings of the 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA)*, 1–3.

40. Su, Y., W. Fan, L. Gao, L. Qiao, Y. Liu, and F. Wu. 2024. "Joint DNN Partition and Resource Allocation Optimization for Energy-Constrained Hierarchical Edge-Cloud Systems." *IEEE Transactions on Vehicular Technology* 72 (6): 8101–12.

41. Sun, N., M. Ding, J. Jiang, et al. 2023. "Cyber Threat Intelligence Mining for Proactive Cybersecurity Defence: A Survey and New Perspectives." *IEEE Communications Surveys & Tutorials* 25 (3): 1748–74. 10.1109/COMST.2023.3283112.

42. Suri, Anshuman, Harsh Chaudhari, Yuefeng Peng, Ali Naseh, Amir Houmansadr, and Alina Oprea. 2026. "Exploiting Leaderboards for Large-Scale Distribution of Malicious Models." *Proceedings of the IEEE Symposium on Security and Privacy (s&p)*.

43. Taleb, T. 2024. "Joint Task and Computing Resource Allocation in Distributed Edge Computing Systems via Multi-Agent Deep Reinforcement Learning." *IEEE Transactions on Network Science and Engineering* 11 (4): 3479–94. 10.1109/TNSE.2024.3479394.

44. Teixeira, P. G., B. G. A. Lebtag, R. P. Dos Santos, J. Fernandes, and A. Mohsin. 2020. "Constituent System Design: A Software Architecture Approach." *Proceedings of the 2020 IEEE International Conference on Software Architecture Companion (ICSA-c)*, 20–27. 10.1109/ICSA-C.2020.00020.

45. Thanvi, Y. S., K. Pappu, and A. Parashar. 2026. "Effect of Shift-Left Security Testing on Early Vulnerability Detection in CI/CD Pipelines." *Proceedings of the IEEE SoutheastCon*, 1–7.

46. Tiwari, Trishita, Ata Turk, Alina Oprea, Katzalin Olcoz, and Ayse K. Coskun. 2019. "User-Profile-Based Analytics for Detecting Cloud Security Breaches." *Proceedings of the ACM Cloud Computing Security Workshop (CCSW)*.

47. Tunc, C. 2023. "Zero Trust Engine for Iot Environments." *Proceedings of the 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA)*, 1–3.

48. Tunc, C. 2025. "A Lightweight and Efficient Intrusion Detection System for Drone Networks." *Proceedings of the 5th Intelligent Cybersecurity Conference (ICSC)*, 219–26.

49. Tunc, C., S. Hariri, and A. Battou. 2019. "A Design Methodology for Developing Resilient Cloud Services (RCS)." In *Handbook of System Safety and Security: Cyber Risk and Management*. Springer. 10.1007/978-3-319-77491-6.

50. Venkatesan, Sridhar, Harshvardhan Sikka, Rauf Izmailov, Ritu Chadha, Alina Oprea, and Michael J. De Lucia. 2021. "Poisoning Attacks and Data Sanitization Mitigations for Machine Learning Models in Network Intrusion Detection Systems." *Proceedings of the IEEE Military Communications Conference (MILCOM)*.

51. Yuan, X. et al. 2024. "Mobility and Cost Aware Inference Accelerating Algorithm for Edge Intelligence." *IEEE Transactions on Mobile Computing* 24 (3): 1530–49.

52. Zhao, D. et al. 2025. "TakeAD: Preference-Based Post-Optimization for End-to-End Autonomous Driving with Expert Takeover Data." *IEEE Robotics and Automation Letters*.

53. Zhao, R., Q. Yuan, J. Li, et al. 2025. "Sce2DriveX: A Generalized MLLM Framework for Scene-to-Drive Learning." *IEEE Robotics and Automation Letters*.