Controlling Artificial Intelligence Risk in Cybersecurity: A Modern Review of Governance, Protection, and the AICURE-Grid Unified Enforcement Framework

Dr. Azhar Hasan Nsaif[1]
bahar299947@uomustansiriyah.edu.iq

Assist. Lec. Uday Kareem Hussein Al-Saadi[2]
udaykareem619@gmail.com

Lec. Nada Abdulkareem Hameed[3]
nada.abdulkarim@muc.edu.iq

Abstract: Artificial intelligence has evolved from a narrow analytical tool into a dual-use cyber capability that supports defensive acceleration while enabling offensive automation. In defense, AI can summarize threat intelligence, triage alerts, detect anomalies, support secure code review, prioritize vulnerabilities, and enrich incident response. The same capabilities may also generate persuasive phishing, automate reconnaissance, mutate malware, exploit leaked credentials, create synthetic impersonation media, and manipulate AI agents through prompt injection or tool misuse. This review treats AI cybersecurity as an adaptive socio-technical risk rather than a conventional software defect and proposes AICURE-Grid, a layered enforcement architecture for controls before, during, and after AI-supported actions. The framework integrates governance, AI asset inventory, identity assurance, an AI gateway, policy decision logic, secure retrieval-augmented generation, data-loss prevention, model and prompt controls, sandboxed tool execution, telemetry, red-team evaluation, and incident response. It argues that model-level safety instructions are necessary but insufficient for high-impact cyber decisions, which require deterministic external controls. A measurable validation plan is provided for security, usability, traceability, and operational performance.

Keywords: Artificial Intelligence Security, Generative AI, LLM Security, Cyber Defense, Risk Management.

[1] Ph.D in Computer Science, Computer Science Department / College of Science / Mustansiriyah University, Iraq.
[2] Assit.Lec. in AI, Ministry of Education, Iraq.
[3] Lec.in Computer Science, Department of Computer Science & Information Technology, Al-Mansour University College, Baghdad, Iraq.

1. Introduction and Research Motivation

Cybersecurity has been reshaped by artificial intelligence in a manner that is deeper than ordinary tool adoption. AI systems are now able to read, classify, generate, summarize, reason, and act over large quantities of security-relevant data. In a security operations center, these abilities can be used to shorten alert triage, accelerate threat-intelligence interpretation, improve vulnerability prioritization, and generate incident documentation. In an adversarial workflow, the same abilities can be used to scale reconnaissance, personalize social engineering, produce convincing multilingual lures, assist malware mutation, and manipulate AI-connected tools. Therefore, the research problem is not whether AI is beneficial or dangerous in isolation. The research problem is how AI can be governed, contained, and measured so that defensive value is obtained while offensive misuse and systemic failure are reduced. The need for a lifecycle control model has been emphasized by modern AI and cybersecurity guidance. The NIST AI Risk Management Framework organizes AI risk management through governance, mapping, measurement, and management [1]. The NIST Cybersecurity Framework 2.0 extends cybersecurity into enterprise governance, risk communication, and resilience outcomes [2]. NIST adversarial machine-learning guidance clarifies that AI systems may be attacked at data, model, algorithmic, deployment, and operational stages [3]. OWASP identifies prompt injection, insecure output handling, training-data poisoning, model denial of service, supply-chain vulnerabilities, sensitive-information disclosure, excessive agency, and systemic risk as major LLM application concerns [4]. Consequently, an AI-cybersecurity proposal must be built around lifecycle control and enforceable architecture, not around prompt wording alone. Secure AI system development and deployment guidance has also made the security-by-design requirement explicit. International guidance led by NCSC, CISA, NSA, and partners recommends that AI systems should be designed, developed, deployed, and operated securely [5]. Additional deployment guidance from NSA, CISA, FBI, and partners emphasizes secure configuration, logging, monitoring, update management, and resilient operation for AI systems [6]. ISO/IEC 42001 introduces requirements for an AI management system, while ISO/IEC 23894 provides guidance for AI-related risk management [7], [8]. The EU AI Act further reinforces the importance of risk classification, documentation, cybersecurity, transparency, human oversight, and quality management for regulated AI systems [9]. These sources collectively show that AI safety must be treated as a governance and assurance program rather than as an

isolated technical filter. A second motivation is derived from software engineering and supply-chain reality. AI applications are not deployed as abstract models only. They are deployed as APIs, RAG pipelines, embeddings, vector databases, plugins, agents, prompt templates, connector permissions, monitoring systems, and continuous delivery workflows. The NIST Secure Software Development Framework and secure-by-design principles are therefore relevant to AI security because vulnerabilities can be introduced through dependencies, insecure code, secrets, misconfiguration, testing gaps, weak logging, or unsafe update practices [10], [11]. MITRE ATLAS provides a threat-informed knowledge base for AI-focused adversarial tactics and techniques, which can be used to build red-team scenarios and measurable controls [12]. A third motivation is operational. Public-sector and industry reports continue to describe a cyber environment shaped by geopolitical tension, cybercrime commercialization, identity abuse, supply-chain interdependence, AI-enabled tactics, and attacker automation [13]-[19]. AI has not removed the importance of cyber fundamentals such as patching, identity assurance, segmentation, logging, backup, secure configuration, and incident response. Instead, AI has increased the speed and personalization of attacks while also improving defensive automation. Figure 1 therefore frames AI as a dual-use capability whose risk or value depends on governance and control.

The contribution of this revised paper is threefold:

1- Recent scientific reviews have been provided for AI-enabled cyber risks, AI-specific attack surfaces, and defensive AI applications.
2- Modern architecture called AICURE-Grid is proposed as a control-oriented model for safe AI-enabled cyber defense.
3- Validation plan is defined so that the architecture can be evaluated by attack success rate, unauthorized tool-call rate, data leakage rate, false blocking rate, time to detect, time to contain, traceability, and analyst productivity.

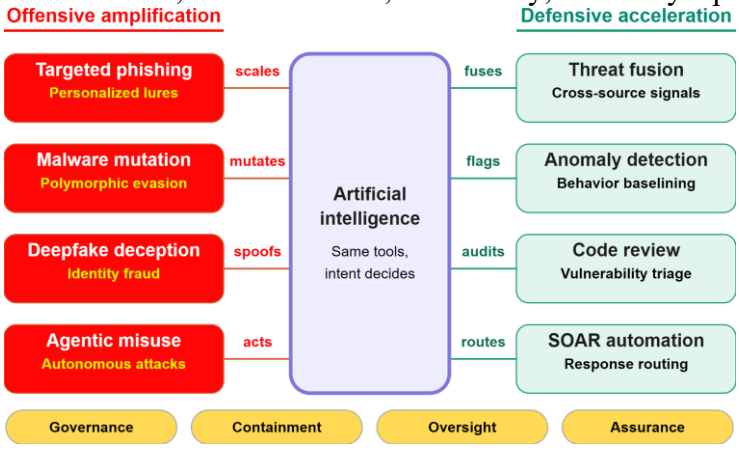

Figure 1: Dual-use position of AI in cybersecurity: offensive amplification and defensive acceleration must be balanced through governance and enforceable controls**.**

2. Methodology and Evidence

The review scope was concentrated on the modern period of generative AI, LLM-based applications, AI agents, and AI governance. Evidence was selected from standards, regulatory sources, government cybersecurity guidance, threat-intelligence reports, systematic reviews, and recent academic works published mainly between 2020 and 2026. The emphasis was placed on sources that can support a practical architecture for organizational deployment, because the research objective is to propose a defensible AI-cyber protection model rather than to describe AI risk at a purely conceptual level. The synthesis was guided by four interrelated analytical dimensions.

1- Attention was directed toward the ways in which artificial intelligence can amplify cybersecurity risks by increasing the speed, scale, personalization, and accessibility of malicious cyber activities.
2- The review examined the distinct attack surfaces introduced by AI systems themselves, including risks associated with data collection, model training, retrieval-augmented generation, inference interfaces, autonomous agents, tool execution, and post-deployment operation.
3- The defensive value of AI was assessed by considering its role in threat intelligence processing, anomaly detection, secure code analysis, vulnerability prioritization, malware investigation, and incident response support.
4- Investigated how a secure and accountable AI-cybersecurity architecture can be designed to combine governance, technical containment, lifecycle assurance, human oversight, and continuous monitoring.

Recent academic surveys were used to validate the breadth of AI-cybersecurity applications and vulnerabilities. A systematic literature review on LLMs and cybersecurity was used to frame cybersecurity-oriented LLM construction, downstream security tasks, and remaining challenges [20]. A broad survey of LLM applications, vulnerabilities, and defenses was used to position network security, software security, cloud security, threat intelligence, critical infrastructure, social engineering, and IoT as affected domains [21]. A systematic review of AI-powered cyber threats was used to support the argument that AI-enabled attacks are not a single technique but an ecosystem of evolving adversarial mechanisms [22]. A

Springer review on generative AI as a double-edged cyber technology was used to support the dual-use framing [23]. A 2026 survey of agentic AI and cybersecurity was included because agentic systems represent a major transition from single-turn generation to persistent action-oriented workflows [24]. The methodology was not limited to academic literature. Standards and operational reports were included because cybersecurity proposals must be implementable within organizations. Standards provide governance structure and control objectives; threat reports provide observed operational pressure; academic research provides technical depth and research gaps. This mapping approach was used to avoid a common weakness in review papers: risk categories are often listed without being connected to deployable controls. Several exclusions were made. Sources that focused only on general AI ethics without cybersecurity implications were not emphasized. Papers that discussed conventional machine learning without connection to modern LLMs, RAG, agentic AI, or secure MLOps were used only where lifecycle security remained relevant. Product marketing claims were not treated as primary evidence unless they reflected a broader operational trend already supported by standards, threat intelligence, or independent research. The result was a focused synthesis suitable for a journal-style review and for a future implementation proposal.

3. AI As Cybersecurity Risk Amplifier

AI increases cybersecurity risk mainly by reducing the cost, time, and expertise required to conduct cyber operations. A traditional attack often requires specialized knowledge across reconnaissance, scripting, social engineering, malware engineering, infrastructure management, and evasion. Generative AI can assist an adversary across several of these stages. It can summarize leaked data, write credible messages, translate lures into many languages, generate code templates, explain vulnerabilities, produce scripts, refine stolen-document analysis, and adapt wording to organizational context. The danger is therefore not limited to wholly new attacks. Existing attacks can be accelerated, scaled, personalized, and made more accessible to lower-skill actors.

1- Social engineering is one of the clearest examples of risk amplification. AI-generated text can be tailored to a victim role, organization, region, event, or internal tone. Deepfake audio or video can make fraudulent requests appear credible, especially when communication is urgent and the victim is accustomed to remote collaboration. Synthetic content can also be combined with compromised identity tokens, exposed organizational charts, and leaked emails. In this combined scenario, traditional user awareness may be weakened because the attack no longer looks generic.

2- Amplification path involves code and malware support. AI tools can generate functional code rapidly, but their security quality can be variable and context dependent. Adversaries may use AI to mutate scripts, modify droppers, document exploit chains, create payload variations, or adapt existing tools to specific environments.
3- Amplification path is agentic autonomy. The model is not only answering; it may be acting. Agent-focused threat modeling has emphasized that reasoning, memory, tool integration, and long-running tasks introduce risks that are not fully addressed by ordinary LLM safety controls [25]. Hybrid prompt-injection research has also shown that prompt manipulation can combine with traditional web weaknesses and multi-agent interactions to create compound attack paths [26].

Prompt injection and jailbreaking are persistent concerns because LLMs are designed to follow instructions expressed in language. Recent reviews of prompt-injection defenses have shown that mitigations include prompt isolation, input filtering, policy classifiers, context separation, output validation, tool governance, adversarial training, and detection layers [27]. However, no single mechanism can be treated as complete protection. Agentic AI attack-surface analysis further indicates that threats may emerge across foundation, cognitive, memory, tool execution, multi-agent coordination, ecosystem, and governance layers [28]. For this reason, the proposed architecture is built as a layered control system. Figure 2 illustrates where lifecycle threats arise.

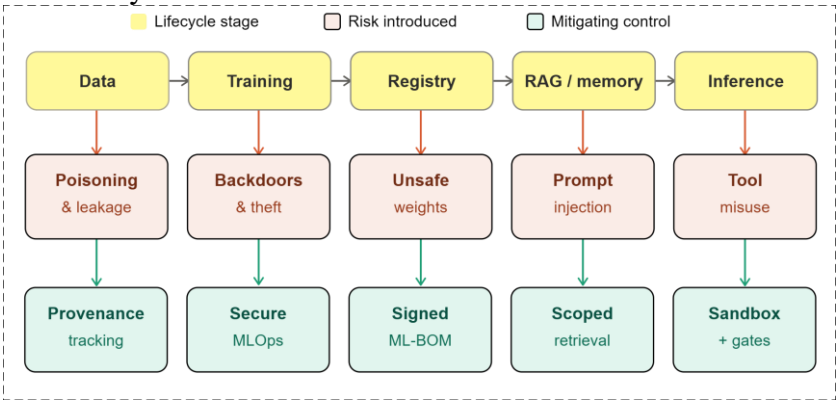

Figure 2: AI lifecycle attack surface: risks must be addressed across data, training, model registry, retrieval, inference, tool use, and operation.

4. AI-Specific Attack and Lifecycle Risk

AI systems create a layered attack surface across the entire lifecycle. During data collection, public sources may be poisoned, labels may be manipulated, licensing may be unclear, and sensitive information may be introduced accidentally. During training and fine-tuning, compromised dependencies, weak access control, insecure experiment tracking, and exposed secrets may allow model theft or behavioral manipulation. During model release, unsigned artifacts, vulnerable containers, and unreviewed adapters may be deployed. During RAG operation, documents and websites may carry indirect instructions that attempt to override system policies. During inference and agent execution, prompt injection, insecure output handling, excessive agency, tool misuse, and privacy leakage become central risks. During operation, drift, missing logs, untested updates, and shadow AI can create long-term exposure. Secure MLOps research has emphasized that the development-to-deployment pipeline must be protected as a unified ecosystem rather than as separate experiments [29]. This is important because a single misconfiguration can affect credentials, data, model artifacts, or production behavior. Adversarial machine-learning surveys have also shown that attacks may occur before, during, and after model training, including poisoning, backdoors, evasion, extraction, and inference-time manipulation [30]. AI risk-management maturity research further indicates that operationalizing responsible AI practices is uneven across organizations, which reinforces the need for measurable maturity stages and evidence-based governance [31]. Therefore, the lifecycle security objective must be broader than front-end content filtering. Controls must cover datasets, labels, notebooks, training code, model registries, containers, prompt templates, embeddings, indexes, APIs, plugins, telemetry, and update policies. Prompt injections must be handled as a trust-boundary problem. Direct prompt injection occurs when a user asks the model to ignore rules, reveal hidden instructions, or perform an unauthorized task. Indirect prompt injection occurs when malicious instructions are embedded in external content, such as an email, web page, PDF, ticket, spreadsheet, code comment, or retrieved knowledge item. The system may fail if the model treats untrusted content as instructions instead of as data. A strong design must therefore separate trusted instructions from untrusted content, label sources, reduce hidden instruction effects, validate outputs, and prevent the model from making unilateral tool-execution decisions. Poisoning and backdoors affect model integrity. In poisoning, an attacker manipulates data used for training, fine-tuning, retrieval, or feedback. In a backdoor, the model behaves normally in most situations but changes behavior when a trigger is present. These threats can be difficult to detect because ordinary evaluation may not include the trigger conditions. In a cybersecurity setting, a poisoned model may misclassify malware,

ignore malicious behavior, recommend unsafe remediation, or disclose sensitive policy details. Dataset provenance, secure labeling, signed artifacts, reproducible training, adversarial test sets, and continuous monitoring are therefore required. Privacy leakage may occur through generated outputs, embeddings, logs, prompts, memory, retrieved context, and tool results. Model extraction attacks may attempt to replicate behavior through repeated queries. Membership inference and inversion attacks may attempt to infer whether specific data was used in training or to reconstruct sensitive information. In enterprise deployments, the most practical exposure often arises from prompt logs, RAG context, long-term memory, and overly permissive connectors. A model connected to internal documents can leak information if access controls are not enforced both at retrieval time and at response-generation time. Table 1 maps lifecycle stages to dominant risks and security objectives. The purpose of this mapping is to show that a control placed only at the user interface cannot protect the entire system. Prompt filtering may reduce direct misuse but cannot prove that retrieved content is trustworthy. Network segmentation may protect an API but cannot prevent a model from generating unsafe instructions. Traditional access control may restrict users but not model-to-tool behavior. For this reason, lifecycle assurance, application security, identity governance, data protection, and operational monitoring must be combined.

Table 1: AI lifecycle risks mapped to security objectives.

Lifecycle stage	Dominant risks	Security objective
Data acquisition	Poisoned data, licensing uncertainty, privacy exposure, biased samples, and weak labeling provenance.	Integrity, provenance, lawful use, minimization, source trust, and dataset documentation.
Training and fine-tuning	Backdoors, compromised libraries, insecure experiments, exposed secrets, model theft, and unreviewed fine-tuning data.	Reproducibility, artifact integrity, access control, secure MLOps, credential protection, and adversarial evaluation.
Model registry and release	Unsafe weights, vulnerable containers, unverified adapters, weak	ML-BOM, signing, dependency scanning, release approval,

	versioning, and unsigned artifacts.	rollback, and controlled deployment.
RAG and memory	Indirect prompt injections, stale documents, unauthorized retrieval, context poisoning, and memory leakage.	Permission-aware retrieval, source trust, memory isolation, document sanitization, freshness, and citation traceability.
Application interface	Jailbreaks, insecure output handling, excessive disclosure, unsafe code generation, and model denial of service.	Instruction control, input constraints, schema validation, output containment, rate limits, and content safety.
Tool execution	Unauthorized email, database, file, cloud, browser, or shell actions by agents.	Least privilege, scoped credentials, PDP approval, sandboxing, reversible execution, and complete audit logging.
Operation and update	Drift, monitoring gaps, untested updates, shadow AI, residual-risk decay, and incident response gaps.	Continuous evaluation, telemetry, red-team regression, model inventory, policy review, and incident playbooks.

5. Positive Contributions of AI to Cybersecurity

Although AI introduces serious risks, it is also one of the most valuable technologies available to cyber defense teams. Modern security operations produce large quantities of alerts, logs, malware samples, vulnerability notices, threat reports, tickets, and incident evidence. Human analysts often face alert fatigue and limited investigation time. AI can be used to summarize evidence, prioritize alerts, correlate weak signals, identify anomalous behavior, draft reports, and recommend remediation steps. When properly governed, AI can increase defensive speed without removing expert judgment. The strongest defensive uses are those in which AI is used for augmentation rather than blind automation. For example, AI can summarize an incident timeline, map observed behaviors to MITRE ATT&CK or ATLAS concepts, draft containment options, and identify related indicators. However, actions that affect production systems, user accounts, legal reporting, external communication, or business continuity should be approved by a human incident commander or by a deterministic policy workflow. This preserves human accountability while reducing the time required for analysis. AI can also improve secure development. It can review code for insecure patterns, explain

vulnerabilities, generate unit tests, suggest safer APIs, detect secrets, and support developer education. However, generated code must be checked because LLM output may be plausible but insecure. Static analysis, dependency scanning, secret detection, unit testing, policy checks, and peer review must remain part of the secure software pipeline. In this model, AI becomes an accelerator of secure engineering rather than a replacement for assurance. Threat intelligence is another strong use case. Security teams must consume large volumes of reports, advisories, indicators, vulnerability disclosures, and incident notes. AI can extract entities, summarize adversary behaviors, cluster related campaigns, generate detection ideas, and support executive reporting. Recent work on AI-focused cyber threat intelligence suggests that conventional CTI processes must be expanded to cover AI-specific assets, artifacts, vulnerabilities, and indicators [32]. Thus, AI can be used not only to defend ordinary systems but also to maintain intelligence about attacks against AI systems themselves.

6. Governance, Standards, and Regulatory Baseline

A scientific proposal for controlling AI risk cannot rely only on technical filtering. Governance is required because AI risk emerges from organizational decisions: which models are used, which datasets are connected, which tools are available, which users can request actions, how outputs are reviewed, how incidents are handled, and how model updates are approved. Technical controls enforce decisions, but governance defines what should be enforced. For this reason, an AI cybersecurity program should contain ownership, risk appetite, policies, roles, inventories, documentation, evidence, escalation paths, and continual improvement. NIST AI RMF and NIST CSF 2.0 can be combined as complementary governance frameworks. The AI RMF contributes AI-specific functions for governing, mapping, measuring, and managing risk [1]. CSF 2.0 contributes enterprise cyber outcomes for governance, identification, protection, detection, response, and recovery [2]. ISO/IEC 42001 adds an AI management-system perspective, and ISO/IEC 23894 provides AI risk-management guidance [7], [8]. The EU AI Act adds legal risk classification, documentation, human oversight, and cybersecurity obligations for relevant systems [9]. Together, these frameworks support the management layer of the proposed architecture.

AI-specific threat frameworks must also be mapped to technical controls. OWASP LLM Top 10 highlights LLM application risks that can be translated into gateway, RAG, output validation, and tool-governance controls [4]. MITRE ATLAS can be used to design AI-focused red-team exercises, detection logic, and incident

response playbooks [12]. Secure-by-design and SSDF guidance can be applied to AI application development, dependency handling, vulnerability disclosure, and secure release practices [10], [11]. The most important governance principle is that AI systems must be inventoried and classified before deployment. Shadow AI is dangerous because it bypasses data classification, legal review, vendor review, logging, monitoring, and incident response. Every AI application should have an owner, purpose statement, model or system card, data-flow diagram, model source, deployment boundary, allowed tool list, risk classification, access policy, evaluation record, logging plan, and retirement procedure. Maturity-model research based on the NIST AI RMF indicates that organizations often struggle to operate AI risk management consistently [31]. This weakness makes governance visibility the first practical control. Governance program must also distinguish between low-risk assistance and high-risk action. A model that summarizes public documentation presents different risk than a security agent that can disable accounts, send emails, deploy code, query production databases, or alter firewall rules. The AICURE-Grid architecture therefore classifies use cases according to data sensitivity, action impact, reversibility, external exposure, and confidence. Decisions are then enforced by the AI gateway and policy decision point rather than by the model alone.

7. Proposed AICURE-Grid Architecture

The proposed architecture is named AICURE-Grid: Artificial Intelligence Cyber Unified Risk Enforcement Grid. It is built on the principle that AI action must be controlled outside the model, not only inside the prompt. The model can be instructed to behave safely, but enforcement of security-sensitive decisions should be performed by deterministic policy components, identity systems, data controls, sandboxed execution environments, and audit mechanisms. This reduces dependence on model obedience and makes the system more suitable for high-stakes cybersecurity contexts. The architecture contains eight interacting layers. The governance layer defines acceptable use, ownership, risk appetite, and approval authority. The AI asset and ML-BOM layer records models, datasets, prompts, embeddings, vector stores, adapters, connectors, and service accounts. The identity and context verify users, services, agents, and non-human identities. The AI gateway mediates prompts, retrieved context, outputs, and tool requests. The policy decision point evaluates action sensitivity, data classification, reversibility, user role, source trust, and model confidence. The secure RAG layer enforces permission-aware retrieval and context isolation. The tool sandbox layer executes allowed actions with least privilege and reversibility. The telemetry and response layer sends evidence to SIEM/SOAR, supports investigation, and enables rollback.

Figure 3 presents the proposed architecture. The figure is deliberately arranged so that the model is not the only center of control. Users and applications interact with an AI gateway, while a policy decision point controls whether model requests and tool calls are permitted. The model may generate reasoning and proposed actions, but it does not directly control enterprise tools. Identity, secure RAG, data protection, tool sandboxing, human approval, audit telemetry, and incident response are placed around the model to create layered containment. The components and expected security outcomes. The architecture can be adapted to multiple deployment models. In a local LLM deployment, the model endpoint may be hosted on-premises and controlled by internal infrastructure. In cloud LLM deployment, the gateway can enforce data minimization, redaction, encryption, and contractual usage limits before prompts are sent externally. In an RAG system, the secure RAG layer can classify and filter retrieved documents. In a SOC copilot, the tool sandbox can restrict actions to case enrichment and draft recommendations unless approval is provided. In a coding assistant, generated code can be passed through static analysis and peer review before merging. In an agentic workflow, the PDP can block or require approval for external communication, destructive changes, and privileged operations. AICURE-Grid is also designed to preserve defensive usefulness. Overly restrictive controls can make AI unusable and may drive employees toward shadow AI. The architecture therefore uses risk-based decision logic instead of a universal deny policy. Low-risk read-only analysis may be allowed with logging. Medium-risk operations may be executed in a sandbox or validated by deterministic checks. High-risk operations may require human approval, be restricted to a change-management workflow, or be denied. This gradation allows AI to improve productivity while preserving accountability.

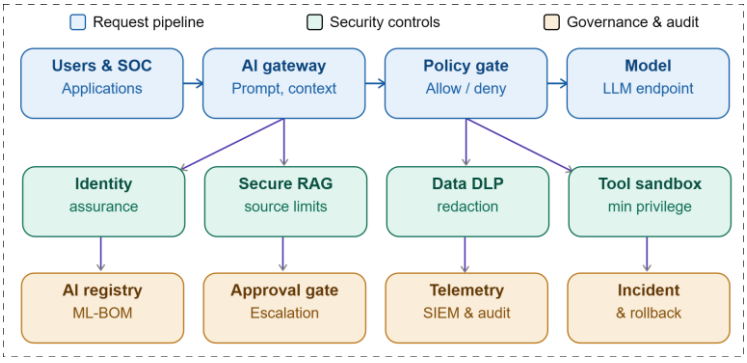

Figure 3: Proposed AICURE-Grid architecture with gateway enforcement, policy decision logic, secure RAG, tool sandboxing, telemetry, and governance oversight.

8. Control Logic, Secure RAG, and Agentic Tool Governance

The policy decision workflow is shown in Figure 4. A model may request action, but the request is enriched with identity, role, data sensitivity, source trust, tool type, target system, reversibility, confidence, and business impact. Low-risk read-only requests can be allowed with logging. Medium-risk requests can be sandboxed, validated, or rate-limited. High-risk requests, such as deletion, external email, access-right changes, production deployment, shell execution, financial operation, or personal-data export, should be approved by a human or denied by default. Agentic penetration-testing research indicates that agent frameworks and models can differ widely in refusal and unsafe-action behavior, which reinforces the need for external enforcement rather than model trust alone [33]. Secure RAG is required because enterprise knowledge retrieval can become a data-leakage or prompt-injection channel. Figure 5 illustrates the secure RAG boundary. Documents should be classified before indexing, access control should be enforced before retrieval, source trust should be scored, hidden instructions should be treated as untrusted content, citations should be preserved, and retrieved context should be labeled. Unknown PDFs, emails, and web pages should not be allowed to override system instructions. Official policies, signed documents, and approved knowledge-based articles may receive higher trust, but permission and freshness must still be checked. Tool governance is the most important difference between a chatbot and an AI agent. A chatbot can produce unsafe text, but an agent can act. It may send email, query a database, update a ticket, run code, modify cloud resources, delete files, or deploy infrastructure. The proposed control principle is that the model may propose an action, but it should not be the final authority for action execution. A deterministic controller must decide whether action is allowed, which credentials can be used, whether a sandbox is required, whether approval is needed, and whether rollback is possible. Recent work on OWASP LLM Top 10 mitigation using intelligent agents suggests that agent-based defenses can be used to detect, assess, and counter LLM risks in real time [34]. However, evaluation work on OWASP defense coverage also shows that defenses may be brittle under paraphrasing and that different defense families close different risk categories [35]. Therefore, architecture should avoid single-control dependency. A refusal-only layer, a keyword filter, or a basic prompt rule is insufficient. Instead, refusal, budget controls, rate limits, tool-registry authentication, credential scrubbing, schema validation, secure RAG, and human approval should be combined. Legal and

governance implications must also be considered for AI agents. Recent regulatory analysis has argued that agent providers face challenges related to inventory of external actions, data flows, connected systems, human oversight, transparency, and runtime behavioral drift under EU-related obligations [36]. Compliance-oriented research connected to the EU AI Act has similarly emphasized cybersecurity, explainability, traceability, failure intervention, and knowledge-base protection for autonomous systems [37], [38]. These observations support the AICURE-Grid emphasis on inventory, traceability, tool-control, and human approval.

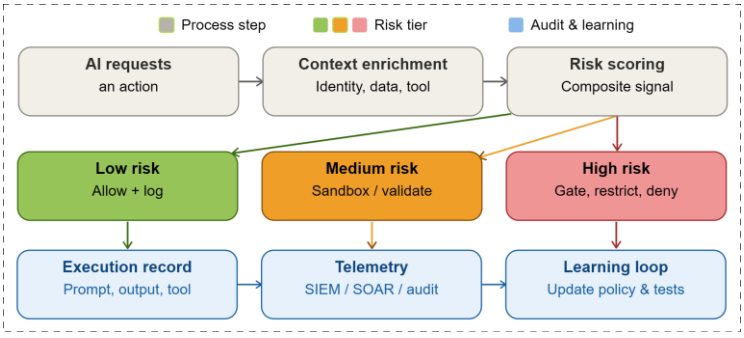

Figure 4: Policy decision workflow for AI tool requests: low-risk actions may be logged, medium-risk actions may be validated or sandboxed, and high-risk actions require approval or denial.

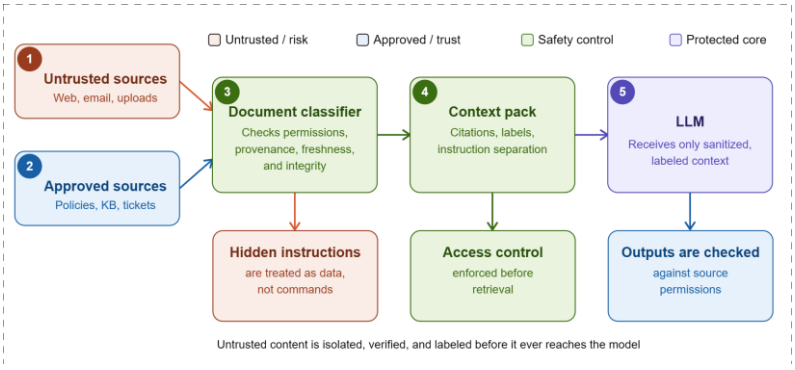

Figure 5: Secure RAG and context isolation boundary: untrusted material is treated as data, permissions are enforced before retrieval, and outputs are checked against source rights.

9. Discussion and Critical Analysis

The AI behavior changes with prompt wording, model version, temperature, retrieved context, memory, tool availability, and policy configuration. Therefore, assurance cannot be treated as a one-time certification. Every model update, retrieval source change, tool integration, policy modification, or new dataset may alter the risk profile. The continuous governance loop shown in Figure 6 is therefore required. Another important observation is that AI risk is often organizational before it becomes technical. Many failures begin with ungoverned adoption, unclear ownership, poor access control, missing logs, weak vendor review, or lack of employee training. A technically advanced model firewall cannot compensate for an organization that does not know which AI systems are used, what data is sent to them, who can authorize tool actions, or how incidents should be escalated. Governance visibility is therefore the first practical defense. The relationship between AI and cyber offense is also nuanced. AI lowers barriers and increases scale, but attackers still need objectives, infrastructure, credentials, exploitable systems, and operational security. Therefore, AI does not eliminate the value of cyber fundamentals. Patch management, strong authentication, segmentation, logging, secure configuration, secrets management, backup, and incident response remain decisive. AI should assist with scale and speed, while human experts, deterministic validators, and policy systems should provide final authority. This balanced model is especially important in cybersecurity because errors can affect confidentiality, integrity, availability, legal obligations, and public trust. Synthetic media and influence operations add a further dimension. Research on synthetic image generation for cyber influence indicates that generative AI can support deception, influence, and subversion scenarios, even when perfect photorealism is not always achieved [39]. This suggests that cyber defense must be connected to identity verification, content provenance, communications governance, and organizational trust procedures. AI risk should therefore be addressed not only in technical SOC processes but also in executive decision-making, communications, HR, finance, and public relations workflows. Secure development literature also confirms that technical and non-technical practices must be combined across the software lifecycle [40].

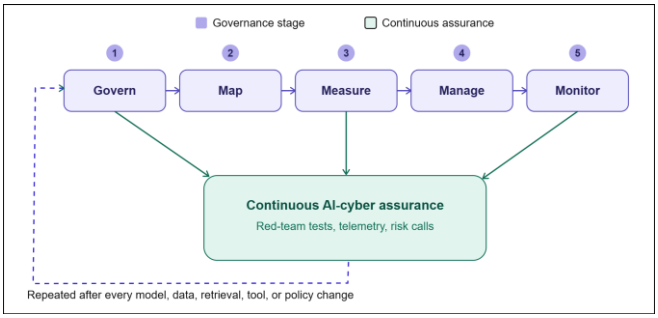

Figure 6: Continuous governance and assurance loop for maintaining AI security after deployment and after every model, data, retrieval, tool, or policy change.

10. Validation Framework and Controlled Augmentation Recommendations

The aim is to determine whether the proposed architecture can reduce AI-enabled abuse while preserving useful defensive capability. Validation should cover security, usability, and operational effectiveness. Security should be measured by the reduction of attack success, prompt-injection bypass, unauthorized tool execution, and data leakage. Usability should be assessed through false blocking, task completion time, and analyst acceptance. Operational value should be evaluated through triage acceleration, incident documentation quality, traceability, and productivity gain. Since recent evidence shows that autonomous AI agents can perform effectively in cybersecurity tasks, evaluation must also verify that such capability is governed, contained, and auditable [41]. AICURE-Grid should therefore be evaluated as a controlled augmentation model. Excessive blocking may reduce operational value, while excessive automation may increase risk.

11. Conclusion

AI should be treated as a dual-use cybersecurity capability. It can support adversaries through social engineering, reconnaissance, code generation, malware adaptation, and agent manipulation, while also strengthening defense through threat intelligence analysis, anomaly detection, alert triage, secure code review, vulnerability prioritization, and incident response. The main recommendation is that AI should be controlled through enforceable architecture rather than model instructions alone. A secure AI-cybersecurity system should combine an AI gateway, policy decision point, secure RAG, data-loss prevention, sandboxed tools, least-privilege access, structured telemetry, red-team testing, and human approval for high-impact actions. Consequently, AICURE-Grid is recommended as a controlled augmentation approach. AI should be used to improve cybersecurity

speed, scalability, and analytical coverage, whereas critical decisions should remain governed by deterministic controls and human oversight. This balance allows defensive benefits to be gained while reducing the risks of uncontrolled autonomy, data leakage, unsafe code generation, prompt-injection compromise, and adversarial misuse.

12. References

[1] E. Tabassi et al., Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1, National Institute of Standards and Technology, 2023. [Online]. Available: https://www.nist.gov/itl/ai-risk-management-framework

[2] National Institute of Standards and Technology, The NIST Cybersecurity Framework (CSF) 2.0, NIST CSWP 29, 2024. [Online]. Available: https://www.nist.gov/cyberframework

[3] A. Vassilev, A. Oprea, A. Fordyce, and H. Anderson, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations, NIST AI 100-2e2023, 2024. [Online]. Available: https://doi.org/10.6028/NIST.AI.100-2e2023

[4] OWASP Foundation, OWASP Top 10 for Large Language Model Applications, 2025. [Online]. Available: https://owasp.org/www-project-top-10-for-large-language-model-applications/

[5] National Cyber Security Centre, CISA, NSA, FBI, and international partners, Guidelines for Secure AI System Development, 2023. [Online]. Available: https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development

[6] NSA, CISA, FBI, ASD, CCCS, NCSC-NZ, and NCSC-UK, Deploying AI Systems Securely: Best Practices for Deploying Secure and Resilient AI Systems, 2024. [Online]. Available: https://www.nsa.gov/Press-Room/Press-Releases-Statements/Press-Release-View/Article/3741371/

[7] ISO/IEC, ISO/IEC 42001:2023, Information Technology--Artificial Intelligence--Management System, International Organization for Standardization, 2023. [Online]. Available: https://www.iso.org/standard/42001

[8] ISO/IEC, ISO/IEC 23894:2023, Artificial Intelligence--Guidance on Risk Management, International Organization for Standardization, 2023. [Online]. Available: https://www.iso.org/standard/77304.html

[9] European Union, Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), 2024. [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[10] M. Souppaya, K. Scarfone, and D. Dodson, Secure Software Development Framework (SSDF) Version 1.1, NIST SP 800-218, 2022. [Online]. Available: https://csrc.nist.gov/pubs/sp/800/218/final

[11] CISA and international partners, Shifting the Balance of Cybersecurity Risk: Principles and Approaches for Secure by Design Software, 2023. [Online]. Available: https://www.cisa.gov/securebydesign
[12] MITRE, ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems, 2025. [Online]. Available: https://atlas.mitre.org/
[13] CISA, Roadmap for Artificial Intelligence, Cybersecurity and Infrastructure Security Agency, 2023. [Online]. Available: https://www.cisa.gov/resources-tools/resources/roadmap-ai
[14] OWASP Gen AI Security Project, Agentic AI--Threats and Mitigations, 2026. [Online]. Available: https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/
[15] World Economic Forum, Global Cybersecurity Outlook 2025, 2025. [Online]. Available: https://www.weforum.org/publications/global-cybersecurity-outlook-2025/
[16] IBM Security, Cost of a Data Breach Report 2025, IBM, 2025. [Online]. Available: https://www.ibm.com/reports/data-breach
[17] Microsoft, Microsoft Digital Defense Report 2025, Microsoft Corporation, 2025. [Online]. Available: https://www.microsoft.com/en-us/corporate-responsibility/cybersecurity/microsoft-digital-defense-report-2025/
[18] Google Cloud Security and Mandiant, M-Trends 2026 Report: Real-World Investigations and Actionable Defense Insights, Google Cloud, 2026. [Online]. Available: https://cloud.google.com/security/resources/m-trends
[19] CrowdStrike, 2026 Global Threat Report, CrowdStrike, 2026. [Online]. Available: https://www.crowdstrike.com/en-us/global-threat-report/
[20] J. Zhang et al., 'When LLMs meet cybersecurity: a systematic literature review,' Cybersecurity, SpringerOpen, 2025. [Online]. Available: https://link.springer.com/article/10.1186/s42400-025-00361-w
[21] N. O. Jaffal, M. Alkhanafseh, and D. Mohaisen, 'Large Language Models in Cybersecurity: A Survey of Applications, Vulnerabilities, and Defense Techniques,' Applied Sciences, vol. 15, no. 17, 2025. [Online]. Available: https://www.mdpi.com/2673-2688/6/9/216
[22] M. Alanezi, 'AI-Powered Cyber Threats: A Systematic Review,' Mesopotamian Journal of Cybersecurity, 2024. [Online]. Available: https://mesopotamian.press/journals/index.php/CyberSecurity/article/view/648
[23] W. Ibrar et al., 'Generative AI: a double-edged sword in the cyber threat landscape,' Artificial Intelligence Review, Springer, 2025. [Online]. Available: https://link.springer.com/article/10.1007/s10462-025-11285-9

[24] M. Gupta and E. Bertino, 'A Survey of Agentic AI and Cybersecurity: Challenges, Opportunities, and Use-case Prototypes,' arXiv:2601.05293, 2026. [Online]. Available: https://arxiv.org/abs/2601.05293
[25] V. S. Narajala and O. Narayan, 'Securing Agentic AI: A Comprehensive Threat Model and Mitigation Framework for Generative AI Agents,' arXiv:2504.19956, 2025. [Online]. Available: https://arxiv.org/abs/2504.19956
[26] J. McHugh, K. Sekrst, and J. Cefalu, 'Prompt Injection 2.0: Hybrid AI Threats,' arXiv:2507.13169, 2025. [Online]. Available: https://arxiv.org/abs/2507.13169
[27] P. H. B. Correia et al., 'A Systematic Literature Review on LLM Defenses Against Prompt Injection and Jailbreaking: Expanding NIST Taxonomy,' arXiv:2601.22240, 2026. [Online]. Available: https://arxiv.org/abs/2601.22240
[28] K. Chu, 'A Systematic Survey of Security Threats and Defenses in LLM-Based AI Agents: A Layered Attack Surface Framework,' arXiv:2604.23338, 2026. [Online]. Available: https://arxiv.org/abs/2604.23338
[29] R. Patel et al., 'Towards Secure MLOps: Surveying Attacks, Mitigation Strategies, and Research Challenges,' arXiv:2506.02032, 2025. [Online]. Available: https://arxiv.org/abs/2506.02032

[30] B. Wu, Z. Zhu, L. Liu, Q. Liu, Z. He, and S. Lyu, 'Attacks in Adversarial Machine Learning: A Systematic Survey from the Life-cycle Perspective,' arXiv:2302.09457, 2023. [Online]. Available: https://arxiv.org/abs/2302.09457
[31] R. Dotan, B. Blili-Hamelin, R. Madhavan, J. Matthews, and J. Scarpino, 'Evolving AI Risk Management: A Maturity Model based on the NIST AI Risk Management Framework,' arXiv:2401.15229, 2024. [Online]. Available: https://arxiv.org/abs/2401.15229
[32] N. Krawczyk, M. Szczepkowski, A. Brodzik, and K. Bocianiak, 'Cyber Threat Intelligence for Artificial Intelligence Systems,' arXiv:2603.05068, 2026. [Online]. Available: https://arxiv.org/abs/2603.05068
[33] V. K. Nguyen and M. I. Husain, 'Penetration Testing of Agentic AI: A Comparative Security Analysis Across Models and Frameworks,' arXiv:2512.14860, 2025. [Online]. Available: https://arxiv.org/abs/2512.14860
[34] M. Fasha, F. Abul Rub, N. Matar, B. Sowan, and M. Al Khaldy, 'Mitigating the OWASP Top 10 For Large Language Models Applications using Intelligent Agents,' arXiv:2601.18105, 2026. [Online]. Available: https://arxiv.org/abs/2601.18105
[35] A. C. Maiorano, 'Which Defense Closes Which Threat? Attributing OWASP-LLM-Top-10 Coverage and Its Brittleness Under Paraphrasing,' arXiv:2606.02822, 2026. [Online]. Available: https://arxiv.org/abs/2606.02822

[36] L. Nannini et al., 'AI Agents Under EU Law,' arXiv:2604.04604, 2026. [Online]. Available: https://arxiv.org/abs/2604.04604
[37] Y. Pita Lorenzo, 'Cumplimiento del Reglamento (UE) 2024/1689 en robótica y sistemas autónomos: una revisión sistemática de la literatura,' arXiv:2509.05380, 2025. [Online]. Available: https://arxiv.org/abs/2509.05380
[38] F. J. Rodriguez Lera et al., 'Aportes para el cumplimiento del Reglamento (UE) 2024/1689 en robótica y sistemas autónomos,' arXiv:2503.17730, 2025. [Online]. Available: https://arxiv.org/abs/2503.17730
[39] M. Mathys, M. Willi, M. Graber, and R. Meier, 'Synthetic Image Generation in Cyber Influence Operations: An Emergent Threat?' arXiv:2403.12207, 2024. [Online]. Available: https://arxiv.org/abs/2403.12207
[40] A. Kudriavtseva and O. Gadyatskaya, 'Secure Software Development Methodologies: A Multivocal Literature Review,' arXiv:2211.16987, 2022. [Online]. Available: https://arxiv.org/abs/2211.16987
[41] V. Mayoral-Vilches et al., 'Cybersecurity AI: The World's Top AI Agent for Security Capture-the-Flag (CTF),' arXiv:2512.02654, 2025. [Online]. Available: https://arxiv.org/abs/2512.02654

التحكم في مخاطر الذكاء الاصطناعي في الأمن السيبراني: مراجعة حديثة للحوكمة والحماية وإطار الإنفاذ الموحّد AICURE-Grid

م.م . عدي كريم حسين السعدي[2]
udaykareem619@gmail.com

م . د . أزهار حسن نصيف[1]
bahar299947@uomustansiriyah.edu.iq

م . ندى عبد الكريم حميد[3]
nada.abdulkarim@muc.edu.iq

المستخلص: تطوّر الذكاء الاصطناعي من أداة تحليلية محدودة النطاق إلى قدرة سيبرانية مزدوجة الاستخدام، تسهم في تسريع العمليات الدفاعية، وفي الوقت نفسه تتيح أتمتة الأنشطة الهجومية. ففي الجانب الدفاعي، يمكن للذكاء الاصطناعي تلخيص معلومات التهديدات، وفرز التنبيهات، واكتشاف السلوكيات الشاذة، ودعم مراجعة الشيفرات البرمجية الآمنة، وتحديد أولويات الثغرات، وتعزيز الاستجابة للحوادث. إلا أن القدرات نفسها قد تُستخدم أيضاً في إنشاء رسائل تصيّد مقنعة، وأتمتة عمليات الاستطلاع، وتعديل البرمجيات الخبيثة بصورة متكررة، واستغلال بيانات الاعتماد المسرّبة، وإنشاء وسائط انتحال اصطناعية، والتلاعب بوكلاء الذكاء الاصطناعي من خلال حقن الأوامر أو إساءة استخدام الأدوات. تتعامل هذه المراجعة مع الأمن السيبراني المرتبط بالذكاء الاصطناعي بوصفه خطراً اجتماعياً وتقنياً متكيفاً، وليس مجرد خلل برمجي تقليدي، وتقترح إطار AICURE-Gridبوصفه بنية إنفاذ متعددة الطبقات لتطبيق الضوابط قبل الإجراءات المدعومة بالذكاء الاصطناعي وأثناءها وبعدها. ويدمج الإطار الحوكمة، وجرد أصول الذكاء الاصطناعي، وضمان الهوية، وبوابة للذكاء الاصطناعي، ومنطق اتخاذ القرارات المستند إلى السياسات، والتوليد المعزز بالاسترجاع الآمن، ومنع تسرب البيانات، وضوابط النماذج والموجّهات، والتنفيذ المعزول للأدوات، وجمع بيانات القياس والمراقبة، واختبارات الفرق الحمراء، والاستجابة للحوادث. وتؤكد الدراسة أن تعليمات السلامة المضمّنة على مستوى النموذج ضرورية، لكنها غير كافية لاتخاذ القرارات السيبرانية عالية التأثير، إذ تتطلب هذه القرارات ضوابط خارجية حتمية وقابلة للإنفاذ. كما تقدّم الدراسة خطة تحقق قابلة للقياس لتقييم الأمن، وسهولة الاستخدام، وقابلية التتبع، والأداء التشغيلي.

الكلمات المفتاحية: أمن الذكاء الاصطناعي، الذكاء الاصطناعي التوليدي، أمن النماذج اللغوية الكبيرة، الدفاع السيبراني، إدارة المخاطر.

---

[1] Ph.D. in Computer Science, Computer Science Department / College of Science / Mustansiriyah University, Iraq.
[2] Asst. Lec. in AI, Ministry of Education, Iraq.
[3] Lec. in Computer Science, Department of Computer Science & Information Technology, Al-Mansour University College, Baghdad, Iraq.