الهيئة الوطنية للأمن السيبراني

National Cybersecurity Authority

Draft of AI Cybersecurity Guidelines

(AICG - 1: 2026)

TLP: Clear

Document Classification: Public

Disclaimer: Please refer to the National Cybersecurity Authority's website (https://nca.gov.sa), to obtain the latest version of this document

In the Name of Allah,
The Most Gracious,
The Most Merciful

Disclaimer: The guidelines in this document have been developed based on Artificial intelligence (AI) cybersecurity best practices, and it is for awareness purposes, in order to reduce and mitigate cybersecurity risks related to AI systems. NCA assumes no responsibility or liability for direct or indirect results of any actions taken based on the information contained in this document. In addition, any contradictions found between this document and the laws and regulation, the entity shall be subjected to the related laws and regulation, NCA strongly recommends each entity to regularly conduct their own assessments to those risks.

Traffic Light Protocol (TLP):

This protocol is widely used around the world and comprises four colors (light signals):

Red (Personal and Confidential- for the Intended Recipient Only)

The recipient is not authorized to share the red-classified material with any individual, whether inside or outside the entity; sharing is strictly prohibited beyond the defined scope of receipt.

Amber+ Strict (Internal Sharing within the Same Entity)

The recipient may share the information only with the intended recipients inside the entity.

Amber (Restricted Sharing)

The recipient may share the information only with the intended recipients inside the entity or with recipients who are required to take action related to the shared information.

Green (Sharing within the Same Community)

The recipient may share the information with others within the same entity or with entities that have a relevant relationship or operate within the same sector; however, it is not permitted to disseminate or publish the information via public channels.

Clear (No Restrictions)

Table of Contents

List of the Figures and Illustrations

List of Tables

Executive Summary

Pursuant to its statute issued by Royal Order No. (6801), dated 11/02/1439H (31/10/2017), the National Cybersecurity Authority (NCA) is the competent authority nationally in charge of cybersecurity in the Kingdom, And the national reference in its affairs. The NCA powers and duties include, but not limited to, the development of cybersecurity policies, governance mechanisms, frameworks, standards, controls, and guidelines; to support the important role of cybersecurity which has increased with the rise of cybersecurity risks in cyberspace more than any time before.

Artificial intelligence (AI) technologies are evolving at a rapid pace, giving rise to cybersecurity vulnerabilities and risks that may differ in nature from traditional cybersecurity risks. With the continued advancement of AI, such as generative AI and agentic AI; cybersecurity considerations can sometimes lag behind innovation. This evolving landscape gives rise to new and complex cybersecurity risks, necessitating dedicated guidelines tailored specifically to AI systems, integrated not only during the design and development phases, but sustained throughout the entire lifecycle, including deployment, ongoing use and retirement.

In this context, the AI Cybersecurity Guidelines (AICG-1: 2026) have been developed after conducting a comprehensive study of multiple international related cybersecurity guidelines, standards, frameworks, and controls, analyzing current status of national initiatives, statistics, and regulatory requirements; to address the unique cybersecurity risk profile of AI technologies faced by entities.

Introduction

The National Cybersecurity Authority (referred to in this document as "The Authority" or "NCA") developed the AI Cybersecurity Guidelines (AICG - 1: 2026) after conducting a comprehensive study of multiple national and international cybersecurity frameworks, standards and controls, and reviewing common industry practices and experiences in the field of AI and AI cybersecurity.

The AI Cybersecurity Guidelines consist of the following:

- 4 Main Domains.
- 15 Subdomains.
- 42 Guidelines.

Objectives

The AI Cybersecurity Guidelines (AICG - 1: 2026) has been developed to support the national cybersecurity objectives by addressing common cybersecurity risks that are unique to AI systems.

The guidelines are defined to address AI cybersecurity risks across the AI system lifecycle, including design and development, deployment, production/operation, and retirement, and to ensure that AI cybersecurity requirements are implemented.

The scope of AICG includes emerging AI technologies such as generative AI and agentic AI, recognizing their distinctive threat vectors and the need for tailored guidelines. Overall this document is designed to set the cybersecurity requirements (that has been previously defined in NCA other published documents), in the context of AI systems specifically, thereby supporting their effective implementation and enhancing cybersecurity across entities.

Figure (1) illustrates the main objectives that AICG aims to achieve, which is to emphasize the importance of protecting the different components of AI Systems as well as protecting against threats arising from the use of AI Systems.

Figure 1: The Main Objectives of AICG

Scope of Work and Applicability

NCA advices every entity in the Kingdom that adopts or plans to adopt to use AI systems to follow these recommended guidelines and implement the minimum cybersecurity best practices in order to minimize cybersecurity risk resulting from the use of AI technology.

These guidelines are not mandatory. However, they are intended to help maintain and enhance the cybersecurity status of the entity while using and adopting AI systems, with the aim of ensuring the application of best practices to reduce cybersecurity risks and to increase resilience from cybersecurity attack.

Due to the dynamic nature of cybersecurity threats; the NCA encourages entities to periodically review and assess cybersecurity risks to determine the need to take additional measures regarding AI systems.

Overview of AI and AI Cybersecurity Threats

Definition of AI Systems

According to The Saudi Data & AI Authority (SDAIA), **AI Systems** are "Software systems that rely on advanced technologies to predict, generate content, provide recommendations, or make decisions, with varying levels of autonomy, depending on the data and the context in which they operate". [1]

Definition of Generative AI

According to The Saudi Data & AI Authority (SDAIA), **Generative AI** is "A branch of artificial intelligence that relies on machine learning techniques and deep neural networks to simulate the human ability to produce original data and content".[2]

Definition of Agentic AI

According to The Saudi Data & AI Authority (SDAIA), **Agentic AI** is "a software system based on AI algorithms, characterized by key features such as perceiving the surrounding environment, interpreting information, setting goals, and making decisions autonomously without constant human supervision".[3]

Components of AI Systems

AI systems consist of a set of distinct interconnected components that ingest data, process inputs, learn patterns, and generate outputs. Generative AI systems are built upon a set of core components that support content generation and reasoning. In some cases, these systems may incorporate additional agentic capabilities that enable planning, tool usage and multi-step task

[1] SDAIA, AI Adoption Framework, Available at: https://sdaia.gov.sa/en/SDAIA/about/Files/AIAdoptionFramework.pdf.
[2] SDAIA, Artificial Intelligence, Available at: https://sdaia.gov.sa/en/SDAIA/about/Pages/AboutAI.aspx
[3] SDAIA, Artificial Intelligence, Available at: https://sdaia.gov.sa/en/SDAIA/about/Pages/AboutAI.aspx

execution to achieve specific objectives. Core Generative AI components may include –but not limited to:

1. **AI Model Components**

   The component responsible for processing input, performing inference, and generating outputs. This may include large language model (LLM), multimodal models, or other AI models supporting the intended AI capabilities.

2. **Data and Knowledge Repositories Components**

   The data repositories and knowledge sources utilized by the AI systems, including enterprise documents, databases, knowledge bases, and other internal or external data sources.

3. **Retrieval and Context Management Components**

   This component is responsible for retrieving relevant information and providing contextual data to the AI model to improve response quality, accuracy and relevance. These capabilities may be implemented through retrieval mechanisms including Retrieval-Augmented Generation (RAG) architectures.

4. **Application and interaction layer Components**

   The interfaces through which users and external systems interact with the AI system, including chat interfaces, user prompts virtual assistant and APIs.

AI systems incorporating agentic capabilities may include additional component that enable autonomous or semi-autonomous execution of task beyond content generation. Examples of such components are- but not limited to:

5. **Planning Components**

   This component is responsible for decomposing objectives into smaller tasks, determining execution sequences and coordinating task completion.

6. **Memory Components**

This component maintains contextual information, historical interaction, intermediate result and task related information to support continuity and decision making through task execution.

7. Tool Integration Components

This component enables interaction with external tools, applications, services, databases, APIs, enterprise systems and standardized integration protocols to support task execution and information retrieval, this may include components such as Model Context Protocol (MCP) server, API gateways, connectors, plugins and service adapters that allow AI models or agents to securely access approved tools and data sources.

Figure (2) illustrates some of the components of Generative and Agentic AI systems and the need for their components to interact outside of AI systems.

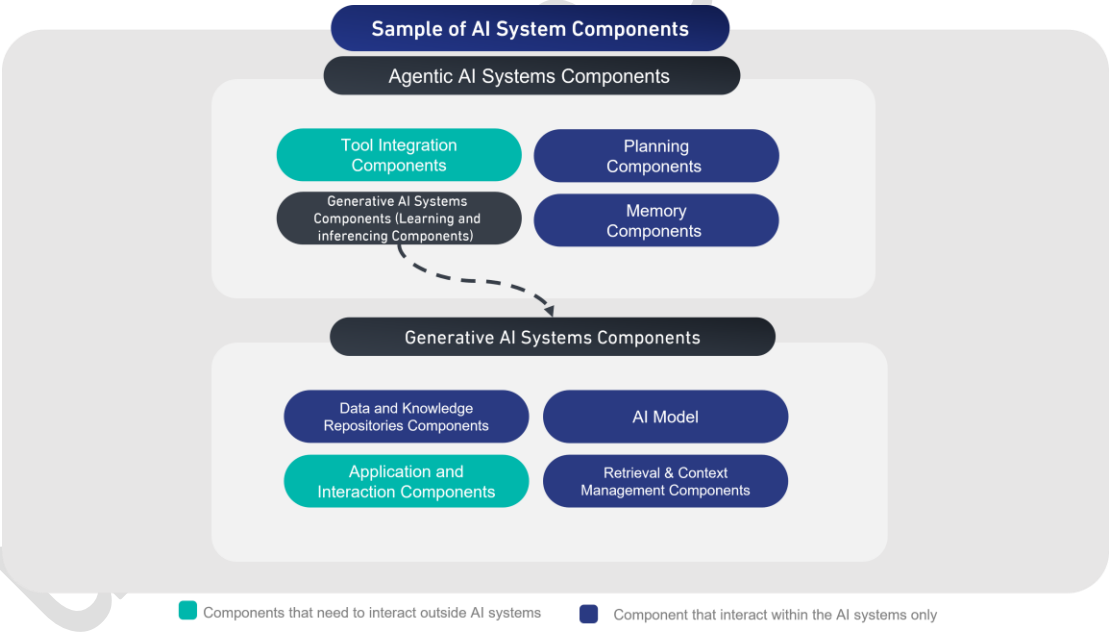

Figure 2: Some of the Components of Generative and Agentic AI Systems

Overview of Major AI Cybersecurity Threats

As using artificial intelligence (AI) brings transformative capabilities to entities, it also introduces a set of cybersecurity threats that must be managed and mitigated. As AI systems rely on large volumes of data, complex algorithms, and interconnected compute environments,

they create a new set of specialized attack surfaces that can be exploited by attackers. This new set of specialized attack surfaces includes data poisoning, model inversion, evasion attacks, and supply-chain compromises of training datasets or model repositories. Because AI decisions often drive critical processes, any manipulation or leakage of the underlying models can result in misinformation, unauthorized actions, or loss of confidentiality, integrity or availability of sensitive information. Effective threat mitigation therefore requires a holistic approach that includes secure data handling, robust model governance, continuous monitoring for anomalous behavior, rigorous testing against adversarial attacks, in alignment with established cybersecurity controls such as the Essential Cybersecurity Controls (ECC). Figure (3) showcases a multitude of attack threats that might be used to target AI systems.

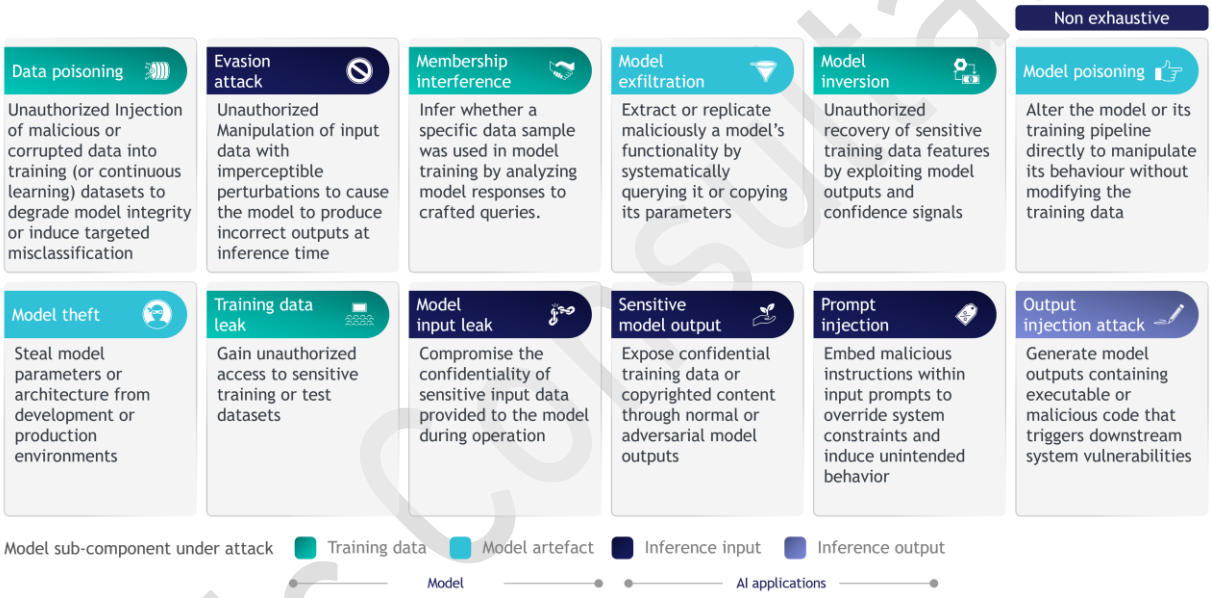

Figure 3: AI Cybersecurity Threats

AI Cybersecurity Guidelines Domains and Structure

Main Domains and Subdomains

Table (1) below shows the Main Domains and Subdomains of AI cybersecurity Guidelines.

1. Cybersecurity Governance	1-1	Cybersecurity Risk Management	1-2	Cybersecurity in Information and Technology Project Management
	1-3	Cybersecurity in Human Resources	1-4	Cybersecurity Awareness and Training Program
2. Cybersecurity Defense	2-1	Asset Management	2-2	Identity and Access Management
	2-3	Network Security Management	2-4	Data and Information Protection
	2-5	Backup and Recovery Management	2-6	Vulnerability Management
	2-7	Penetration Testing	2-8	Cybersecurity Event Logs and Monitoring Management
	2-9	Web Application Security		
3. Cybersecurity Resilience	3-1	Cybersecurity Resilience Aspects of Business Continuity Management (BCM)		
4. Third-Party Cybersecurity	4-1	Third-Party Cybersecurity		

Table 1: Main Domains and Subdomains of AI Cybersecurity Guidelines

AICG Documentation Structure

The AICG itself is referred to as described in Figure (4) and Figure (5).

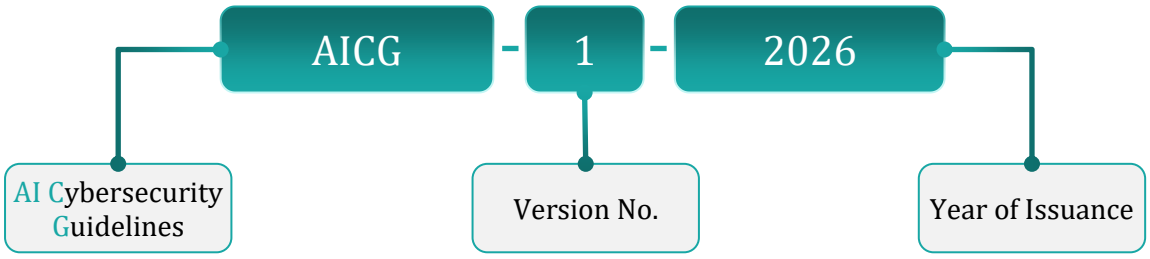

Figure 4: AICG Coding Scheme

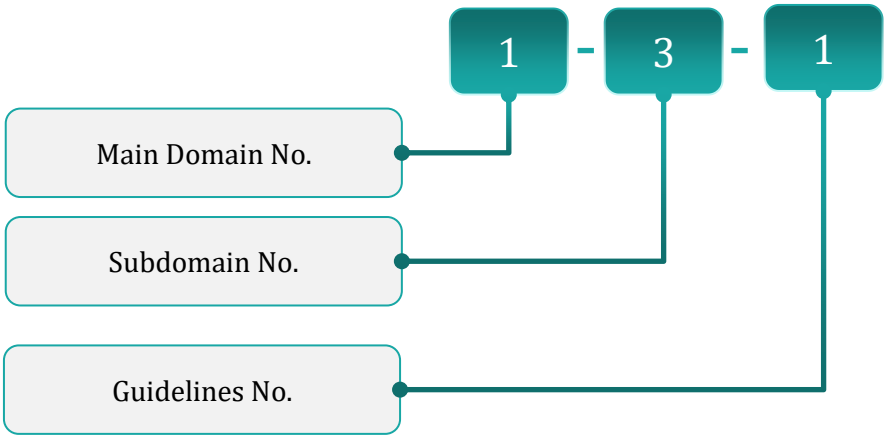

Figure 5: AICG Structure

Table (2) below shows the methodological structure of the Guidelines.

1 → 	Name of Main Domain
Reference number of the Main Domain	Name of Subdomain
Objective	
Guidelines	
Guidelines Reference Number	Guidelines Clauses

Table 2: AICG Structure

AI Cybersecurity Guidelines

1 Cybersecurity Governance

1-1	Cybersecurity Risk Management
Objective	To ensure that AI cybersecurity risks are managed in a methodological approach in order to protect the entity's information and technology assets as per organizational policies and procedures, and related laws and regulations.
Guideline	
1-1-1	Define, document, approve and implement cybersecurity risk management practices across all the lifecycle of AI systems.
1-1-2	Conduct a cybersecurity risk assessment to classify AI agents' actions by potential impact, likelihood and reversibility, and apply appropriate safeguards.
1-2	**Cybersecurity in Information and Technology Project Management**
Objective	To ensure that AI cybersecurity requirements are included in information and technology project management practices throughout the lifecycle of AI systems.
Guideline	
1-2-1	Implement security by design across the AI systems lifecycle, along with defining and implementing procedures to secure AI development and operation, from the definition of cybersecurity requirements and secure data acquisition, through model training, fine-tuning, prompt configuration, evaluation, deployment, operation and maintenance to decommissioning.
1-2-2	Treat the data retrieved from external tools, memory bases, or third-party APIs as untrusted, and apply strict validation and sanitization to any data injected into the context window to prevent the execution of malicious commands disguised as data.
1-2-3	Avoid model retraining or uncontrolled learning directly in production, unless explicitly justified with cybersecurity risk assessed and protected by specific controls.
1-2-4	Continuously asses the cybersecurity capabilities and resilience of AI agents throughout the development lifecycle.

1-2-5	Conduct a systematic cybersecurity code review, cybersecurity testing and formal approval of all AI-generated code to address cybersecurity risks before use in development, integration or production environments.
1-2-6	Establish and enforce approval processes for high-impact AI prior to production deployment, ensuring cybersecurity risks are assessed, addressed.
1-2-7	Implement a graduated autonomy to increase AI agent independence, ensuring continuous human oversight and oversight to monitor and understand cybersecurity-related decisions and actions.
1-3	**Cybersecurity in Human Resources**
Objective	To ensure that AI cybersecurity risks and requirements related to personnel are managed efficiently prior to employment, during employment and after termination/separation as per organizational policies and procedures, and related laws and regulations.
Guidelines	
1-3-1	Conduct screening or vetting on personnel who have access to internal AI systems components or who have privileged access on production AI environments, or other AI functions that have high risks.
1-3-2	Ensure confidentiality, acceptable use, and secure handling obligations for all personnel whose roles may expose sensitive data, prompts, outputs, or configurations through AI systems.
1-4	**Cybersecurity Awareness and Training Program**
Objective	To ensure that personnel and users receive AI-specific cybersecurity awareness and training commensurate with their roles and the cybersecurity risks associated with AI systems.
Guidelines	
1-4-1	Include AI cybersecurity training for personnel involved in AI system development, configuration, or operation, including: • Secure prompt design, tool design, output handling, and cybersecurity incident reporting. • Secure handling of sensitive data in prompts and uploads, recognition of AI-related cybersecurity threats related to AI (such as prompt injection or data leakage), and

	validation of outputs, in addition to the awareness of relevant cybersecurity measures to protect from the potential security risks.
1-4-2	Include AI system cybersecurity awareness within the entity overall cybersecurity awareness program, including: • Secure use of prompt and user provided input data when interacting and the AI system, including the avoidance of sensitive information in prompt and inputs which can cause cybersecurity risks such as data leakage. • Cybersecurity risks associated with AI-generated code and the required review and testing controls before use.

2 Cybersecurity Defense

2-1	Asset Management
Objective	To ensure that the entity have an accurate and detailed inventory of AI-related information and technology assets in order to support cybersecurity and operational requirements and maintain the confidentiality, integrity and availability of those assets.
Guidelines	
2-1-1	Establish and continuously update an inventory of assets for all AI system components.
2-1-2	Define asset owners and lifecycle status for assets for all AI system components, including experimental, pilot, pre-production, approved production and restricted and retired assets, services or configurations.
2-1-3	Define and implement procedures for the assets and components after using AI systems and during decommissioning such as the secure disposal for AI systems.
2-2	**Identity and Access Management**
Objective	To ensure the secure and restricted logical access to AI-related information and technology assets in order to prevent unauthorized access and allow only authorized access for users, services and AI agents necessary to accomplish assigned tasks.
Guidelines	
2-2-1	Define a role-based permissions and authorization policies for human users, service accounts, and AI agents while ensure applying identity and access control principles (Need-to Know and Need-to-Use principle, Least Privilege principle, and Segregation of Duties principle).
2-2-2	Implement periodic review of identities, access rights, tokens, and service accounts used by AI systems.
2-2-3	Ensure the ability to control authorizations in order to mitigate AI cybersecurity risks including rate limiting, usage threshold and other mechanisms to mitigate model extraction and data leakage risks.
2-3	**Network Security Management**

Objective	To ensure the cybersecurity of network paths, interfaces and communications used by AI systems, including training, tuning, retrieval and inference traffic, and communications with external tools and services.
Guidelines	
2-3-1	Segment the network internal AI system components from external components, not explicitly required for their operation and restricting protocols, ports, destinations and egress paths to approved requirements only, in order to enforce trust boundaries and detect anomalies.
2-3-2	Segment the network external AI system components and restrict the access and the communications, only as needed for operational purposes, while applying appropriate controls to protect internet-facing systems, reduce the attack surface and detect unusual situations.
2-4	**Data and Information Protection**
Objective	To ensure the protection of data and information used by, generated by or otherwise associated with AI systems throughout their lifecycle.
Guidelines	
2-4-1	Protect the confidentiality and integrity of AI system components data, including training data, evaluation data, model weights, prompts, embeddings, logs, contexts and outputs at rest, in transit and during processing, and restrict access and disclosure to authorized personnel only.
2-4-2	Protect the data classified in AI systems at rest, in transit and during processing, based on its classification in accordance to the relevant laws and regulations requirements.
2-4-3	Protect AI memory components from unauthorized modification and validate stored content to prevent memory poisoning and manipulation.
2-4-4	Develop guardrails to detect and prevent malicious input, unauthorized actions, sensitive information disclosure and any other cybersecurity threats and ensure testing before production use and after material changes.
2-5	**Backup and Recovery Management**

Objective	To ensure secure backup, recovery and rollback arrangements for AI systems.
Guidelines	
2-5-1	Implement secure backup and recovery mechanisms to ensure availability, integrity, and a timely restoration of AI systems.
2-5-2	Test backup, recovery and rollback procedures for AI systems periodically.
2-6	**Vulnerability Management**
Objective	To ensure that vulnerabilities affecting AI models, applications, infrastructure, dependencies and integrations are identified, assessed, remediated and communicated in a timely manner.
Guidelines	
2-6-1	Identify, track and remediate vulnerabilities in AI systems
2-6-2	Maintain subscriptions or equivalent access to relevant threat and vulnerability intelligence affecting AI systems, and monitor developer security advisories, service status updates, cybersecurity alerts.
2-6-3	Disable or restrict vulnerable AI features, until remediation is validated, when a vulnerability is detected.
2-7	**Penetration Testing**
Objective	To assess the effectiveness of cybersecurity guidelines protecting AI systems by simulating relevant attack techniques and threat scenarios, in order to identify exploitable cybersecurity vulnerabilities, misconfigurations, and weaknesses prior to deployment and periodically thereafter, commensurate with their cybersecurity risk, complexity, and exposure.
Guidelines	
2-7-1	Conduct AI system components and AI system specific penetration testing activities before deployment and periodically thereafter, including testing outputs and system behavior.
2-7-2	Conduct AI system testing through different methods such as red teaming and adversarial evaluation and validate human override.

2-7-3	Address the discovered findings and vulnerabilities by cybersecurity testing and evaluation and define the proper remediation procedures.
2-8	**Cybersecurity Event Logs and Monitoring Management**
Objective	To ensure that logging, monitoring and detection capabilities cover AI-specific events, actions, changes and anomalies needed to support compliance, cybersecurity incident response and continuous assurance.
Guidelines	
2-8-1	Log and monitor cybersecurity-relevant AI components logs, including authentication and access events, privileged actions, dataset and knowledge source access, prompt configuration changes, plugin and tool invocations.
2-8-2	Log and monitor actions taken by AI systems and cybersecurity-relevant events, including data uploads and downloads, and AI-initiated actions that may impact the confidentiality, integrity or availability.
2-8-3	Protect, retain and analyze cybersecurity-relevant AI-related logs to support compliance, cybersecurity incident response, investigations and vulnerability remediation.
2-9	**Web Application Security**
Objective	To ensure that AI-enabled applications, APIs, web interfaces, agents, plugins and related integrations are designed, developed, configured and operated securely.
Guidelines	
2-9-1	Protect AI systems web applications and services using secure communication protocols, request rate limiting, denial-of-service protections, controlled egress points, and filters to block unsafe channels.

3 Cybersecurity Resilience

3-1	Cybersecurity Resilience Aspects of Business Continuity Management (BCM)
Objective	To ensure that AI systems are addressed within business continuity and disaster recovery arrangements in a manner commensurate with their importance and dependencies.
Guidelines	
3-1-1	Define a rollback, safe shutdown and manual alternatives for AI systems in the event of a cybersecurity incident, where applicable.

4 Third-Party Cybersecurity

4-1	Third-Party Cybersecurity
Objective	To ensure that cybersecurity risks arising from third-party entities, components and services used in the AI lifecycle are identified, assessed, managed and monitored on an ongoing basis.
Guidelines	
4-1-1	Assess and Manage cybersecurity risks that may arises from the use of AI systems by third parties and entities
4-1-2	Assess and manage the cybersecurity of third-party related AI third party components in AI systems components before use and on an ongoing basis.
4-1-3	Ensure AI Developers' compliance with the entity AI cybersecurity requirements is contractually established, and monitoring and periodically assessing such compliance.
4-1-4	Define exit, portability and secure disposal requirements for AI system components upon developer change or contract termination.

Annexes

Annex No. (A): Terminologies and Definitions

Table 3 below shows some of the terminologies contained herein, and the meanings as described thereto.

Term	Definition
AI Agent	A software entity or capability that can perceive input, invoke tools or external functions, and act with some degree of autonomy to achieve specified objectives.
AI System	According The Saudi Data & AI Authority (SDAIA), AI systems are "Software systems that rely on advanced technologies to predict, generate content, provide recommendations, or make decisions, with varying levels of autonomy, depending on the data and the context in which they operate". This term includes AI enabled services and all the AI system components (generative and agentic AI systems) defined in this document such as AI Models Components, Data and Knowledge Repositories Components, Retrieval and Context Management Component, Application and interaction layer Components, Planning Components, Memory Components and Tool Integration Components.
Agentic AI	According to The Saudi Data & AI Authority (SDAIA), **Agentic AI** is "a software system based on AI algorithms, characterized by key features such as perceiving the surrounding environment, interpreting information, setting goals, and making decisions autonomously without constant human supervision".
Generative AI	According to The Saudi Data & AI Authority (SDAIA), **Generative AI** is "A branch of artificial intelligence that relies on machine learning techniques and deep neural networks to simulate the human ability to produce original data and content".
Embeddings	Numerical representations of data used by AI systems for retrieval, similarity matching or other model-driven operations.
Fine-tuning	Additional training or adaptation of a pre-existing AI model using task-specific or entity-specific data.
Inference	The process by which an AI model produces an output from input data, prompts or other context.

Term	Definition
Model Weights	The learned parameters of an AI model that influence its behavior and outputs.
Prompt	Input text, instructions or context provided to an AI model to influence its response or action.
Prompt Injection	An attack technique that manipulates prompts, context or retrieved content to alter model behavior, bypass controls or induce unauthorized actions.
Training Data	Data used to train, pre-train, fine-tune or otherwise adapt an AI model.
Tool / Plugin	An external function, connector, service or interface that an AI system can invoke or use during operation.

Table 3: Terms and Definitions

Annex No. (B): List of the Abbreviations

Table 4 below shows the abbreviations used in this document.

Abbreviation	Meaning
AI	Artificial Intelligence
AICG	AI Cybersecurity Guidelines
API	Application Programming Interface
BCM	Business Continuity Management
ECC	Essential Cybersecurity Controls
LLM	Large Language Model
MCP	Model Context Protocol
NCA	National Cybersecurity Authority
RAG	Retrieval Augmented Generation
TLP	Traffic Light Protocol

Table 4: List of Abbreviations


الهيئة الوطنية
للأمن السيبراني
National Cybersecurity Authority