ISO/IEC JTC 1/SC 42

ISO/IEC AWI TS 42119-7(en)

Secretariat: ANSI

Artificial intelligence — Testing of AI —

Part 7:
Red teaming

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: + 41 22 749 01 11
E-mail: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Foreword .... 오류! 책갈피가 정의되어 있지 않습니다.

Introduction .... 오류! 책갈피가 정의되어 있지 않습니다.

Scope .... 오류! 책갈피가 정의되어 있지 않습니다.

Normative references .... 오류! 책갈피가 정의되어 있지 않습니다.

Terms and definitions .... 오류! 책갈피가 정의되어 있지 않습니다.
Terms and definitions .... 오류! 책갈피가 정의되어 있지 않습니다.
Abbreviations .... 오류! 책갈피가 정의되어 있지 않습니다.

Introduction to AI red teaming for AI systems .... 오류! 책갈피가 정의되어 있지 않습니다.
General .... 오류! 책갈피가 정의되어 있지 않습니다.
Comparison between traditional red teaming and AI red teaming 오류! 책갈피가 정의되어 있지 않습니다.

Methodologies for planning and executing AI red team assessments 오류! 책갈피가 정의되어 있지 않습니다.
Three-phase approach to AI red teaming .... 오류! 책갈피가 정의되어 있지 않습니다.
Phase 1: team formation and preparation .... 오류! 책갈피가 정의되어 있지 않습니다.
Phase 2: execution of AI red team activities .... 오류! 책갈피가 정의되어 있지 않습니다.
Phase 3: knowledge sharing and reporting .... 오류! 책갈피가 정의되어 있지 않습니다.

Red teaming techniques .... 오류! 책갈피가 정의되어 있지 않습니다.
Overview of AI red teaming techniques .... 오류! 책갈피가 정의되어 있지 않습니다.
Selecting appropriate red teaming techniques .... 오류! 책갈피가 정의되어 있지 않습니다.
Selection of testing approach based on system components .. 오류! 책갈피가 정의되어 있지 않습니다.

Procedures for documenting and reporting findings 오류! 책갈피가 정의되어 있지 않습니다.
Data recording and management .... 오류! 책갈피가 정의되어 있지 않습니다.
Report structure .... 오류! 책갈피가 정의되어 있지 않습니다.
Communication Protocols .... 오류! 책갈피가 정의되어 있지 않습니다.
Confidentiality maintenance .... 오류! 책갈피가 정의되어 있지 않습니다.
Differentiation between identification and measurement ...... 오류! 책갈피가 정의되어 있지 않습니다.

Integration into AI development and deployment lifecycle .... 오류! 책갈피가 정의되어 있지 않습니다.

Guidance on adapting traditional software testing methodologies 오류! 책갈피가 정의되어 있지 않습니다.
Evolution of test case design .... 오류! 책갈피가 정의되어 있지 않습니다.
Utilization of automated testing tools .... 오류! 책갈피가 정의되어 있지 않습니다.
Interpretation of results of red teaming .... 오류! 책갈피가 정의되어 있지 않습니다.

(informative) Use case examples .... 오류! 책갈피가 정의되어 있지 않습니다.

(informative) AI red teaming techniques .... 오류! 책갈피가 정의되어 있지 않습니다.

(informative) Examples of document templates .... 오류! 책갈피가 정의되어 있지 않습니다.

(informative) Example factors that can cause previously unknown vulnerabilities to emerge .... 오류! 책갈피가 정의되어 있지 않습니다.

(informative) Relationship with ISO/IEC/IEEE 29119-2오류! 책갈피가 정의되어 있지 않습니다.

Bibliography... 오류! 책갈피가 정의되어 있지 않습니다.

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

ISO draws attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO takes no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO [had/had not] received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents.ISO shall not be held responsible for identifying any or all such patent rights.

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee [or Project Committee] ISO/IEC JTC 1, Information Technology, Subcommittee SC 42. Artificial Intelligence.

This edition cancels and replaces the edition (ISO :), which has been technically revised.

The main changes are as follows:

A list of all parts in the ISO series can be found on the ISO website.

Any feedback or questions on this document should be directed to the user’s national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

Artificial Intelligence (AI) systems have evolved significantly, with large-scale AI systems based on Large Language Models (LLMs) and Foundation Models becoming central to various applications across industries. These models possess unprecedented capabilities, but their complexity and scale introduce unique safety, robustness, and trustworthiness challenges. Concerns about misuse, unintended behaviors, and vulnerabilities in such models necessitate robust evaluation methodologies.

Red teaming has emerged as a critical practice for proactively identifying and mitigating potential risks in AI systems, especially those using LLMs and Foundation Models. While traditional red teaming focuses on evaluating physical and cyber systems through simulated adversary testing, AI red teaming specifically addresses the quality and trustworthiness of AI systems. This standard provides comprehensive guidelines for conducting AI red team assessments, emphasizing large-scale models and incorporating a wide range of red teaming techniques and methodologies.

This document is part of the ISO/IEC 42119 series on testing AI systems and is intended to be used in conjunction with the ISO/IEC 29119 software testing series.

Artificial intelligence — Testing of AI — Part 7:Red teaming

1 Scope

This document provides technology-agnostic guidance for conducting red teaming assessments on AI systems. The document covers:

- Definitions and terminology specific to red teaming in AI contexts
- Identification of risks, applicability, objectives and attack vectors
- Methodologies for planning and executing red teaming assessments on AI systems, aligned with ISO/IEC/IEEE 29119 series processes as appropriate
- Procedures for documenting and reporting red teaming findings
- Recommendations for integrating red teaming into the AI system lifecycle

This document is applicable to testing all types of AI systems.

NOTE PE ACTION-09: we revisit the scope later (with CA1-035-047, **JP041-052, JP040-053**)

We will add the detailed reference information with ISO/IEC 29119 series(Part1, Part2), ISO/IEC 42119 series(part2).

PE ACTION-12: Based on the definition, we will discuss the change of scope text, replace or removing the assessment word.

PE ACTION-13: Based on the definition, we will discuss the change of scope text, what will focusing on AI systems or AI models or both.

PE ACTION-14: Based on the definition, we will discuss the change of scope text, with provided missing elements.

- Roles and responsibilities;
- Competencies including skills
- Disclosure criteria
- Preferrably also add the reference to Bias and ethic standards.

2 Normative references

There are no normative references in this document.

3 Terms and definitions

For the purposes of this document, the *following terms and definitions / terms and definitions given in , as well as the following [delete what doesn't apply]* apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

- IEC Electropedia: available at http://www.electropedia.org/
- ISO Online browsing platform: available at http://www.iso.org/obp

3.1 Terms and definitions

3.1.1
Red team
independent group that challenges a defined target to improve its effectiveness or robustness by assuming an adversarial role or point of view

EXAMPLE Target can be an organization, system, plan, or product

3.1.2
AI red team
stakeholders with varying backgrounds who specifically focus on identifying vulnerabilities in AI systems through adversarial methods

3.1.3
Adversarial attack
attemp to deceive AI models into making incorrect or unintended predictions or classifications using crafted inputs or data to exploit vulnerabilities

3.1.4
Data Poisoning
Manipulation of training data to maliciously influence the model's behavior

3.1.5
Hallucination
Instances where AI models generate outputs that are nonsensical or factually incorrect, particularly relevant in large scale AI systems

3.1.6
incident
anomalous or unexpected event, set of events, condition, or situation at any time during the life cycle of a project, product, service, or system

[SOURCE: ISO/ IEC/IEEE 29119-1:2022]

3.1.7
test specification
complete documentation of the test design, test cases (3.1.8) and test procedures (3.1.10) for a specific test item

[SOURCE: ISO/ IEC/IEEE 29119-1:2022]

3.1.8
test case
set of preconditions, inputs and expected results (3.1.9), developed to drive the execution of a test item (3.1.11) to meet test objectives (3.1.12)

[SOURCE: ISO/ IEC/IEEE 29119-1:2022]

3.1.9
expected result
observable predicted behaviour of the test item (3.1.11) under specified conditions based on its specification or another source

[SOURCE: ISO/ IEC/IEEE 29119-1:2022]

3.1.10
test procedure
sequence of test cases (3.1.8) in execution order, with any associated actions required to set up preconditions and perform wrap up activities post execution

[SOURCE: ISO/ IEC/IEEE 29119-1:2022]

3.1.11
test item
test object
work product to be tested

EXAMPLE Software component, system, requirements document, design specification, user guide.

Note 1 to entry: Other terms used for "test item" are "test object", "software under test" and "system under test".

[SOURCE: ISO/ IEC/IEEE 29119-1:2022]

3.1.12
test objective
reason for performing testing

EXAMPLE Checking for correct implementation, identification of defects, measuring quality.

[SOURCE: ISO/ IEC/IEEE 29119-1:2022]

3.1.13
test plan
detailed description of test objectives (3.1.12) to be achieved and the means and schedule for achieving them, organized to coordinate testing (3.131) activities for some test item (3.1.11) or set of test items

[SOURCE: ISO/ IEC/IEEE 29119-1:2022]

3.1.14
testing
set of activities conducted to facilitate discovery and evaluation of properties of test items (3.1.11)

[SOURCE: ISO/ IEC/IEEE 29119-1:2022]

3.1.15
test
activity in which a system or component is executed under specified conditions, the results are observed or recorded, and an evaluation is made of some aspect of the system or component

[SOURCE: ISO/ IEC/IEEE 29119-1:2022]

3.1.16
red teaming
<artificial intelligence> testing practice focused on inducing an AI model or system to behave in unintended ways

Note 1 to entry: Red teaming can use either a benign or adversarial perspective to simulate a wide range of realistic threats or challenging scenarios or situations.

3.1.17
machine learning algorithm
algorithm to determine *parameters* of a *machine learning model* from data according to given criteria

[SOURCE: **ISO/IEC 22989:2022, 3.3.6**]

3.1.18
Agentic AI system

AI system capable of perceiving its environment, reasoning about how to achieve high-level goals, and executing actions using tools or external interfaces with limited direct human supervision.

Note 1 to entry: Agentic AI systems often include capabilities such as planning, memory management, and tool orchestration.

3.1.19
Multi-agent system

System composed of multiple interacting intelligent agents that can collaborate, coordinate, or compete to solve problems that are difficult or impossible for an individual agent or a monolithic system to solve.

3.1.20
Indirect prompt injection

Attack technique where malicious instructions are embedded in external data sources (e.g., websites, emails, documents) that the AI agent retrieves and processes, causing the agent to execute unintended actions.

3.2 Abbreviations

FM	foundation model
LLM	large language model
MMLM	multimodal language model
VLA	visual language action
VLM	vision language model

4 Introduction to AI red teaming for AI systems

4.1 General

AI Red Teaming simulates actual attacks adopting an attacker's mindset to identify vulnerabilities and threats, and verifies the resilience of the AI system.

Traditional software testing methodologies, which typically focus on verifying compliance with pre-defined functional specifications, are often insufficient for large-scale AI systems such as those based on Large Language Models (LLMs) and Foundation Models. These models possess unprecedented capabilities, but their complexity and scale introduce unique challenges regarding safety, robustness, and trustworthiness that traditional verification cannot fully address.

Specifically, traditional testing may fail to detect:

- Unintended Behaviors: Issues like hallucinations (nonsensical or factually incorrect outputs) or the generation of toxic content, which occur despite correct functional implementation.
- Adversarial Vulnerabilities: Susceptibility to specialized attacks such as prompt injection (direct or indirect) or data poisoning, where crafted inputs manipulate the model into violating safety policies.
- Emergent Risks: Previously unknown vulnerabilities that emerge from real-world interactions, evolving news events, or new cybersecurity research after the model's training phase.

Therefore, AI Red Teaming is essential to proactively identify and mitigate these risks by simulating realistic threats and adversarial interactions that go beyond standard functional verification.

Unlike traditional red teaming, which focuses on organizational processes and infrastructure vulnerabilities, AI red teaming targets inherent weaknesses in AI models and data. It employs adversarial machine learning techniques to uncover factors that lead to reduced accuracy, robustness, or trustworthiness.

The primary objectives of AI red teaming include the following four key areas:

- Verification of Detection and Response Capabilities: Assessing how effectively the defense organization (e.g., Blue Team) detects AI-specific attacks and measuring the speed and efficiency of their response processes (MTTD/MTTR).
- Realistic Threat Simulation: Utilizing Tactics, Techniques, and Procedures (TTPs) similar to those of actual attackers to simulate realistic scenarios.
- Identification of Holistic Blind Spots: Evaluating vulnerabilities across People, Processes, and Technology rather than focusing solely on software bugs.
- Assessment of Business Impact: Visualizing attack paths to critical assets ("Crown Jewels") to demonstrate the actual business consequences.

The scope of the assessment extends beyond the AI model at its core to include the entire AI system as a whole, verifying impacts on downstream components and external integrations. Furthermore, this practice should be an iterative process conducted throughout the system lifecycle—from development to operation—rather than a single event prior to deployment.

This document specifically addresses these methodologies within the framework of ISO/IEC TS 42119-2 and aligns with the dynamic testing processes of the ISO/IEC/IEEE 29119 series.

PE_NOTE: call for expert contribution by IN05010-010 (ISO/IEC 5338 , ISO/IEC/IEEE 16085 , ISO/IEC 20246 , ISO/IEC TS 25058 , ISO/IEC DIS 25059, ISO/IEC 29147 , [8]

4.2 Comparison between traditional red teaming and AI red teaming

This clause provides a brief comparative overview of traditional red teaming and AI red teaming.

— **Objective focus**:

— Traditional Red teaming: Identifies areas for improvement by simulating realistic conditions across physical and digital systems, revealing potential shortcomings and ways to enhance overall performance.

  - AI red teaming: Targets inherent weaknesses in AI and machine learning models, including factors that lead to reduced model accuracy or robustness, and aims to guide enhancements in model design and data handling.

- **Methodology and techniques**:

  - Traditional Red Teaming: Employs scenario-based simulations, examination of human and technical interactions, and comprehensive evaluations of hardware and software functionalities.

  - AI Red Teaming: Uses specialized adversarial machine learning techniques, such as generating adversarial examples, manipulating input data and analyzing model behaviors to highlight and address algorithmic limitations.

- **Expertise required**:

  - Traditional Red Teaming: Requires broad knowledge of system operations, including physical and digital domains, cybersecurity, and an understanding of how various components interact under realistic conditions.

  - AI Red Teaming: Requires deep expertise in machine learning algorithms , data analysis, and model-specific vulnerability identification to refine model architectures and data processing strategies.

- **Attack vectors**:

  - Traditional Red Teaming: Investigates hardware, software and user interactions to uncover potential inefficiencies, design flaws or process gaps that can be remedied for improved system performance.

  - AI Red Teaming: Examines mathematical and algorithmic structures, as well as data flows, to identify where models can fail and how they can be made more robust and accurate.

- **Outcome and remediation**:

  - Traditional Red Teaming: Results in strengthened system stability, more efficient operations and refined workflows through adjustments to processes and components.

  - AI Red Teaming: Leads to enhanced AI model reliability, improved decision-making processes, and the implementation of safeguards that ensure models perform consistently and effectively under a range of conditions.

Table 1 — Comparison table contrasting AI-specific red teaming approaches with traditional and cybersecurity red teaming methods

Aspect	**Traditional Red Teaming**	**Cybersecurity Red Teaming**	**AI system Red Teaming**
Primary Target	Strategies and Plans	IT Systems and Networks	AI System, models and behaviors
Attack Surface	Organizational processes	Security vulnerabilities	AI Data, Model and System’s adversarial vulnerabilities and weaknesses
Expertise Required	Strategy and Domain knowledge	Security and IT skills	AI/ML technology, data, and AI trustworthiness and safety expertise

Tools and Techniques	Workshops and War games	Penetration testing tools	AI evaluation frameworks and dataset, AI testing, frontier AI testing
Outcome Focus	Strategic improvement	Security hardening	AI system safety enhancement

4.3 Multi-dimensional approaches to AI red teaming

AI Red Teaming can be conducted across various dimensions depending on the organization's objectives. The assessment shall consider the following dimensions as appropriate:

- **Security & Safety (including CBRN):** Focuses on identifying high-stakes risks such as the generation of instructions for Chemical, Biological, Radiological, and Nuclear (CBRN) threats, self-harm, or violence. The goal is to verify that the system refuses to generate harmful cont Team selection ent.

- **Quality (Reliability & Robustness):** Focuses on inducing hallucinations, factual inconsistencies, or logical errors through adversarial inputs to measure the model's trustworthiness.

- **Performance (Efficiency under Attack):** Focuses on identifying inputs that cause disproportionate resource consumption (e.g., sponge attacks) or latency degradation, distinct from standard load testing.

4.4 Relationship with other standards

AI Red Teaming interacts closely with existing ISO/IEC standards to ensure a comprehensive evaluation framework:

- **Lifecycle Integration (ISO/IEC 5338):** Red teaming activities are conducted iteratively across the AI system lifecycle processes defined in ISO/IEC 5338, serving as a critical verification method during both development and operation phases.

- **Risk Management (ISO/IEC/IEEE 16085):** This document supports the risk management processes of ISO/IEC/IEEE 16085. Red teaming findings serve as inputs for risk identification and estimation, while also verifying the effectiveness of implemented risk controls.

- **Quality Evaluation (ISO/IEC 25059 & TS 25058):** Red teaming specifically targets the AI-specific quality characteristics defined in ISO/IEC 25059, such as robustness, safety, and trustworthiness. It functions as a dynamic evaluation technique consistent with the guidance in ISO/IEC TS 25058.

- **Process and Reporting (ISO/IEC 29147 & 20246):** Vulnerabilities discovered during red teaming shall be handled in accordance with the disclosure principles of ISO/IEC 29147. Additionally, test plans and reports generated during the process should undergo work product reviews as outlined in ISO/IEC 20246 to ensure accuracy and consensus.

5 Methodologies for planning and executing AI red team assessments

5.1 Three-phase approach to AI red teaming

AI red teaming is a systematic process divided into three phases:

— **Phase 1**: team formation and preparation

— **Phase 2**: execution of AI red team activities

— **Phase 3**: knowledge sharing and reporting

Red teaming can be an input or output from or to risk management, for example, risk identification, risk estimation, risk/threat modeling and risk control verification.

The principal procedures of this standard follow the dynamic test process of ISO/IEC/IEEE 29119-2.

5.2 Phase 1: team formation and preparation

5.2.1 AI red team composition and management

5.2.1.1 Team selection

5.2.1.1.1 Red team

The organization should establish the red team to align with the organization's overall management structure, standards for development and procurement, and business risks, to effectively plan and promote the red teaming exercise. Furthermore, the organization shall appoint subject matter experts (SMEs) based on the specific risk scenarios:

- **For CBRN/Dual-use risks:** Experts in biology, chemistry, or national security are required to validate the feasibility of generated threats.
- **For Quality/Hallucination:** Experts in the specific knowledge domain (e.g., legal, medical) or linguists are required to verify factual accuracy and reasoning.
- **For Performance/Availability:** Engineers capable of monitoring system latency and resource usage under adversarial conditions are required.

5.2.1.1.2 Attack planner/conductor

If internal resources are insufficient, the organization shall engage external security experts or third-party resources.

5.2.1.1.3 Target AI system development and provision manager

The development and provision manager of the target Al system **should** manage the process with consideration of the impact of red teaming on the release schedule, etc.

The organization shall structure the red team based on the scale and characteristics of the organization and the target AI system.

5.2.1.1.4 Other relevant stakeholders

Top management or designated executives shall oversee the red teaming process to ensure alignment with organizational objectives.

The organization should involve risk management stakeholders to address broader business risks beyond information security.

The organization should evaluate potential business risks and impacts during the formulation of risk scenarios and the preparation of the final assessment report.

Personnel with expertise in relevant external interfaces should be assigned to the red team if the target AI system interacts with other information systems.

This should be considered when selecting team members, if necessary.

5.2.1.2 Timing of red teaming

5.2.1.2.1 General

The organization shall determine the scope and timing of red teaming in accordance with the system evaluation schedule.

5.2.1.2.2 Before the release

The red team shall conduct the initial red teaming prior to the release of the target AI system.

The scope of red teaming shall cover the entire target AI system rather than being limited to specific subsystems."

The scope of red teaming, conducting structure, conducting cost, schedule, etc. should be individually planned and conducted according to the status of the target Al system.

5.2.1.2.3 After the release

The organization should conduct periodic red teaming throughout the system development, deployment, and operation phases.

5.2.1.2.4 Training and resources

5.2.1.2.4.1 Competence and Training

The organization shall ensure that red team members possess the necessary skills and knowledge to effectively identify risks in the target AI system. Training programs should be provided or required, covering the following areas:

- **System Knowledge:** Understanding the architecture, intended use, data flow, and safeguards (e.g., guardrails) of the target AI system.
- **Adversarial Techniques:** Training on current AI red teaming methodologies, including prompt injection, jailbreaking, and evasion techniques relevant to the model type (e.g., LLMs, Computer Vision).
- **Domain-Specific Risks:** For specialized assessments (e.g., CBRN, Medical, Finance), training on domain-specific terminologies, regulations, and risk indicators is required to accurately assess the impact of generated outputs.

- **Ethics and Rules of Engagement (RoE):** Clear guidelines on the scope of testing, prohibited actions (e.g., attacking production infrastructure causing DoS), data privacy handling, and responsible disclosure procedures.

5.2.1.2.4.2 Tool and Environment Support

The organization shall provide the red team with adequate resources to simulate realistic attack scenarios and document findings effectively. This includes:

- **Computational Resources:** Sufficient computing power (e.g., GPUs) and budget for API usage to conduct extensive testing without resource bottlenecks.

- **Red Teaming Tools:** Access to automated testing frameworks, fuzzing tools, vulnerability scanners, and logging platforms designed for AI systems.

- **Access Privileges:** Appropriate levels of access to the system (Black-box, Gray-box, or White-box) including documentation, model weights, or system prompts, as defined in the scope.

5.2.1.2.4.3 Tester Safety and Psychological Support

Given the nature of AI red teaming, which often involves exposure to toxic, violent, or disturbing content (e.g., hate speech, CSAM, self-harm descriptions), the organization should prioritize the well-being of the red team members.

- **Psychological Safety:** The organization should provide access to psychological support services and implement rotation schedules to minimize prolonged exposure to harmful content.

- **Opt-out Mechanisms:** Team members should have the right to opt-out of testing specific high-risk categories that may cause personal distress.

5.2.2 Planning the AI red ream assessment

5.2.2.1 Development of risk scenarios

5.2.2.1.1 Identify risk scenarios

The red team shall identify risk scenarios considering the following aspects:

a) System configuration: Identify the architecture of the target AI system, including whether it utilizes commercial services (API-based), open-source software (OSS), or in-house developed models . This includes mapping data flows and interactions between the AI model and other system components.

b) System usage patterns: Analyze how the AI system is utilized, including:

a) Output usage: How the AI outputs (e.g., code generation, SQL queries) are used by downstream systems.

b) Reference sources: Whether the system uses Retrieval-Augmented Generation (RAG) or accesses external internet resources.

c) Model interaction: Potential risks regarding the AI model itself, such as training data poisoning or feedback loop exploitation.

c) Information assets to be protected: Identify critical assets within the system, such as confidential organizational data, personal identifiable information (PII), or intellectual property that could be exposed through the AI system.

d) **Evaluation Dimensions and Perspectives:** Align risk scenarios with specific evaluation dimensions:

- **Safety & Security:** Scenarios involving the generation of **CBRN (Chemical, Biological, Radiological, Nuclear)** material, assisting in cyberattacks, or bypassing safety filters (Jailbreaking).
- **Quality & Integrity:** Scenarios aimed at inducing **hallucinations**, bias, or reasoning failures that degrade the utility of the system.
- **Adversarial Performance:** Scenarios involving inputs designed to exhaust system resources (e.g., increasing token generation time or memory usage) to degrade service availability.
- **Agentic Control & Authorization:** Scenarios involving the agent executing unauthorized actions, escalating privileges (e.g., role inheritance exploitation), or interacting with critical systems without human-in-the-loop verification.
- **Memory & Context Integrity:** Scenarios where the agent's long-term memory or context window is manipulated to alter future behavior or exfiltrate sensitive information across sessions.
- **Systemic & Multi-Agent Risks:** Scenarios involving cascading failures in multi-agent environments, orchestration manipulation, or unlimited resource consumption (e.g., infinite loops).
- **Checker-Out-of-the-Loop**: The risk of bypassing human approval processes or missing notifications during autonomous actions.
- **Critical System Interaction**: The risk of issuing incorrect commands to physical systems or IoT devices.

5.2.2.1.2 Tasking AI red team

The proposal should include the organization of the red team, threats and vulnerabilities in the Al system, purpose and necessity of red teaming, overview of the target system conducting outline, schedule, estimated cost, and proposed structure.

If the conducting of red teaming is included in the organization's risk management procedures, the red teaming should be conducted in accordance with the relevant procedures.

The formation of the red team should include the 'attack planner/conductor' and 'Experts from Relevant Domains' as described.

5.2.2.2 Objective setting

5.2.2.2.1 Defining Primary Goals

The organization shall clearly define the primary goals of the red teaming assessment prior to execution. The objectives should be specific, measurable, achievable, relevant, and time-bound (SMART). The objectives shall align with the organization's risk tolerance and the intended use of the AI system.

5.2.2.2.2 Classification of Objectives

To ensure comprehensive coverage, the objectives should be classified into the following dimensions, as applicable to the system's risk profile:

- **Security Assurance:** To verify the system's resistance to active attacks, such as jailbreaking, prompt injection, or extraction of training data.

- **Safety Verification:** To confirm that the system refuses to generate harmful content, including CBRN (Chemical, Biological, Radiological, Nuclear) instructions, hate speech, or self-harm encouragement.

- **Reliability and Quality Assessment:** To measure the rate of hallucinations (factually incorrect outputs) or logical failures under adversarial conditions.

- **Performance Stability:** To assess the system's latency and resource consumption when subjected to complex or malicious inputs (e.g., sponge attacks) intended to degrade availability.

5.2.2.2.3 Success Criteria and Metrics

The red team shall establish clear criteria for what constitutes a "successful" attack or a "passed" assessment.

- **Quantitative Metrics:** e.g., "The Attack Success Rate (ASR) for jailbreaking attempts must be below 1%," or "Latency under adversarial load must not exceed 200ms."

- **Qualitative Metrics:** e.g., "The system must not reveal any PII (Personally Identifiable Information) regardless of the prompt complexity."

5.2.2.2.4 Alignment with Compliance and Policy

The objectives should also include verifying compliance with:

- Internal AI governance policies (e.g., ethical guidelines).

- External regulatory requirements (e.g., EU AI Act, specific industry safety standards).

5.2.2.3 Scope definition

The scope for the red teaming should include not only the AI models at its core but the entire Al system as a whole.

The attack planner should utilize the Software Bill of Materials (SBOM) or AI Bill of Materials (AIBOM) to identify system components.

If the target Al system includes individual AI systems for specific functions (e.g., search query generation, answer generation, inspection), the attack planner/conductor should classify each AI systems accordingly:

The scope of red teaming activities should be defined by analyzing the system architecture, usage contexts, extensions (e.g., plug-ins), and existing security controls of target AI system.

In such cases, the feasibility of conducting black-box testing and white-box testing for each component should be confirmed. This should be discussed with the parties concerned in advance.

Confirmation levels such as whether to simply indicate the possibility of a successful attack, provide evidence regarding the likelihood of a successful attack, or confirm that the attack will actually succeed, should be set based on the following factors:

Testing activities should exclude resource-exhaustion attacks, such as Denial of Service (DoS), to prevent unintended service disruption.

Therefore, when conducting red teaming, the scope of red teaming should be determined after consultation with the parties concerned, assuming the consequences and damage that may be caused.

5.2.2.4 Resource allocation

The organization shall ensure the provision of necessary resources, including budget, personnel, and organizational structure, to support red teaming activities.

The organization should also identify and allocate other resources such as necessary tools.

External experts should be engaged to fulfill the role of attack planner or conductor if internal resources are insufficient.

5.2.2.5 Target environment preparation

The red team should determine the appropriate environment for testing (e.g., in-operation, staging, or development) to minimize potential damage to live services. When conducting tests, the red team should coordinate with system administrators to adjust monitoring configurations (e.g., temporary allow-listing in IDS/IPS, suppression of specific alerts) to ensure that attack signatures are not blocked prematurely, while maintaining necessary audit logs.

5.2.2.6 Organizing the schedule

The red team should plan their red teaming activities by taking into account the release schedule and development status of the target Al system.

In doing so, guided by the timing concepts described in clause X.X they should also consider segmenting the system into components or layers, and aligning with various test schedules of the system (including unit, integration, and system tests).

The red team should arrange the schedule, taking into account the quality and quantity of risk and attack scenarios to be developed in clause X.X.

5.2.2.7 Preparing the environment for AI red teaming

The red team shall request required assets, including URLs, API keys, and logs, from the system administrator.

The red team shall coordinate with relevant stakeholders to implement temporary modifications, such as alerting suppression, required for testing.

When testing involves third-party services, the red team should adhere to service agreements and coordinate with providers for log acquisition.

The red team should notify relevant stakeholders (involved in systems affected by the execution of the attack scenarios) with the plan of the red teaming, the scope of impact, the schedule, and other relevant details.

Relevant stakeholders shall be informed of the assessment plan, including the scope, schedule, and potential impacts, prior to execution.

5.2.2.8 Legal and ethical considerations

— Objectives: Align techniques with specific goals, such as verifying robustness, detecting biases or upholding widely accepted ethical standards.

5.2.2.9 Escalation and emergency procedures

The red team shall establish a clear escalation flow to address unexpected system behaviors, critical failures, or security incidents arising during the red teaming exercise.

This procedure shall include:

- Stop/Go Criteria: Defined thresholds for suspending the red teaming activity if operational risks exceed acceptable levels
- Test Incident Reporting: Immediate communication channels for reporting critical vulnerabilities discovered prior to the final report

If the organization already has an escalation flow for overall crisis management or security incidents, the red team should follow that process accordingly.

In addition, the red team should agree on the following: the evaluation of the assumed damage and scope of impact, the stop/go criteria for the red teaming exercise, and the remediation procedures for unexpected behavior, failures, or issues.

The escalation flow for such cases should also be confirmed.

5.2.3 Preparation for red teaming

5.2.3.1 Tasking AI red teamers

5.2.3.1.1 Rules of Engagement (RoE)

The attack planner shall establish strict Rules of Engagement (RoE) to ensure safety and legal compliance. The RoE shall define:

- **Forbidden Targets:** Systems or data that must not be touched (e.g., production databases containing real PII, critical infrastructure control interfaces).
- **Authorized Techniques:** Specific approval for high-risk techniques (e.g., generating actual malware code vs. pseudo-code).
- **Stop Conditions:** Criteria for immediately halting testing (e.g., discovery of a vulnerability that allows full system takeover or unintended leakage of real user data).

5.2.3.1.2 Domain-Specific Mission Assignments

Tasks shall be assigned based on the expertise of the red teamers and the specific evaluation dimensions:

- **CBRN & Safety Team:** "Attempt to bypass safety filters to generate actionable instructions for creating hazardous materials or self-harm methods. Focus on 'jailbreaking' prompt sequences."
- **Quality & Integrity Team:** "Focus on 'adversarial fact-checking'. Provide inputs designed to trigger hallucinations or logical inconsistencies in the model's reasoning process."
- **Performance Team:** "Execute 'sponge attacks' or complex query injections to identify thresholds where system latency or resource consumption degrades beyond acceptable levels."

5.2.3.1.3 Reporting Requirements

Red teamers shall be instructed on the required format for reporting findings, which must include:

- The full prompt chain used to trigger the behavior.
- The model's output (screenshot or log).

- **Severity Assessment:** A preliminary rating of the risk (e.g., Critical, High, Medium, Low) based on the domain impact (e.g., "Critical" for successful CBRN generation).

5.2.3.2 Instructions and guidelines

A communication plan, including stakeholder identification, securing communication channels, and feedback and remediation, needs to be established. For related information, refer to Clause 7.3.

5.2.4 Analysis of test item

5.2.4.1 Architectural & Data Flow Analysis

Before execution, the red team shall analyze the target system's architecture to identify potential attack surfaces. This includes:

- **Input/Output Modalities:** Identifying all entry points (text, image, file upload) and exit points.
- **External Dependencies:** Analyzing connected plugins, APIs, or databases (RAG) that could be exploited for indirect prompt injection.
- **Training Data & Knowledge Base (White-box only):** Reviewing the data sources for potential bias, poisoned data, or hazardous information (e.g., presence of chemical synthesis papers in the training set).

5.2.4.2 Review of Existing Safeguards

The red team shall analyze the existing defense mechanisms to plan effective bypass strategies:

- **System Prompts:** Analyzing the "meta-prompt" or system instructions that define the AI's persona and constraints.
- **Filtration Layers:** Identifying input/output filters (e.g., keyword blocking, semantic classifiers) to design evasion techniques (e.g., encoding attacks, hypothetical scenarios).

5.2.4.3 Dimension-Specific Analysis Strategy

The analysis shall focus on specific components relevant to the evaluation goals:

- **For Performance Testing:** Analyze the system's context window limits, timeout settings, and compute resource allocation to design effective resource exhaustion attacks.
- **For Quality Testing:** Analyze the specific domain knowledge base (e.g., medical guidelines document) to craft counterfactual questions that test the model's adherence to ground truth.

5.3 Phase 2: execution of AI red team activities

5.3.1 Basic concept

A tester conducts tests based on the defined test plan and test specifications, which guide and constrain the testing activities. The tests are executed to identify potential weaknesses in the system under test. These weaknesses are fed back, providing valuable feedback for improving the test plan, refining the test specifications.

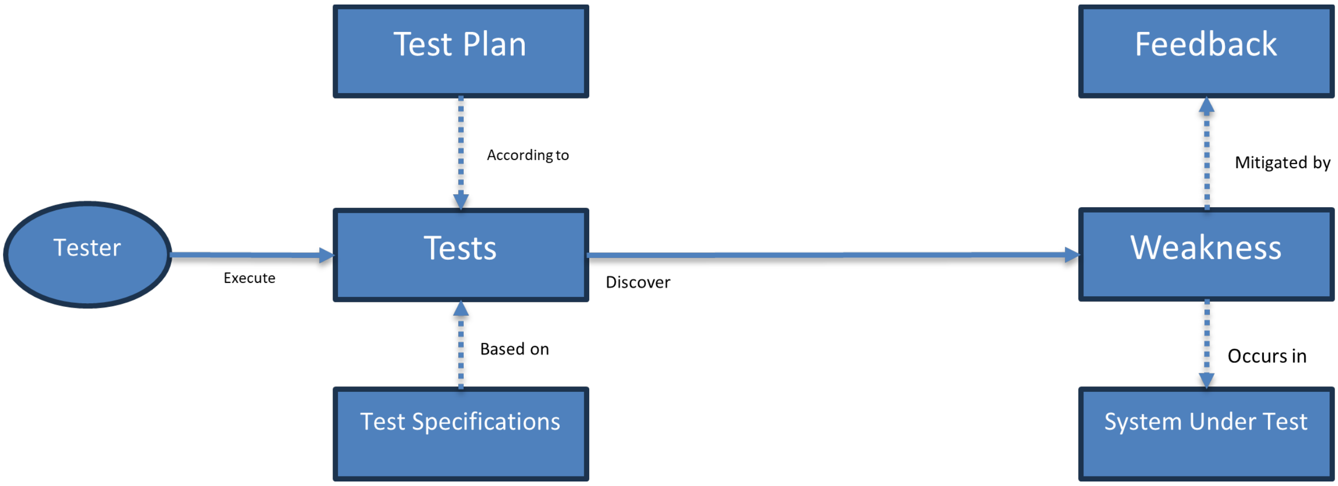

Figure 1 — Relationships among key elements in red teaming

The red team should execute the assessment using the following three-step process:

a) Step 1 (Exploratory Testing): The red team should conduct initial testing on individual prompts to identify effective attack vectors

b) Step 2 (Attack Development): The red team should develop customized attack signatures and payloads tailored to the identified risk scenarios.

c) Step 3 (System-wide Testing): The red team should execute the attack procedures against the full AI system to verify downstream impacts.

5.3.2 Reconnaissance

Red teaming must take into account various risk scenarios based on the relevant guidance.

Prior to developing specific attack vectors, the red team shall identify risk scenarios based on the following four factors :

- System Configuration: Analysis of components (e.g., commercial APIs, OSS, in-house models) and data flows.

- Usage Patterns: Evaluation of how the AI outputs are used (e.g., code generation) and what sources are referenced (e.g., RAG, internet access).

- Information Assets: Identification of critical assets to be protected, such as PII, intellectual property, or internal knowledge bases.

- AI Safety Evaluation Perspectives: Consideration of safety dimensions such as toxicity, fairness, privacy, and security assurance.

Risk scenarios and attack scenarios should be considered for each evaluation perspectives

5.3.3 Identifying and accessing targets

5.3.3.1 Identification of Attack Surfaces

The red team shall identify specific entry points and components based on the evaluation objectives:

- **Interface Endpoints:** Identify all user interaction points, including chat interfaces, API endpoints (REST/gRPC), and file upload features.

- **Data Ingestion Pipelines (for Indirect Injection):** Identify external data sources accessed by the system, such as RAG (Retrieval-Augmented Generation) databases, web search modules, or third-party plugins, which can be targets for data poisoning or indirect prompt injection.

- **Model Components (White-box):** Identify access paths to model weights, embeddings, or system prompts if the scope allows for white-box testing.

5.3.3.2 Target Selection by Risk Domain

Targets shall be prioritized according to the specific risk domain:

- **CBRN & Safety:** Focus on the "System Prompt" or "Safety Filter" layers to attempt bypasses (jailbreaking). Target the RAG knowledge base to check if restricted documents (e.g., dual-use technology papers) are accessible.

- **Performance:** Identify resource-intensive endpoints (e.g., long-context processing, image generation) that are susceptible to sponge attacks or latency degradation.

- **Quality & Reliability:** Target specific domain-knowledge modules (e.g., the medical advice subsystem) to test for hallucinations or reasoning failures.

5.3.4 Attack development

The attack planner/conductor (including third parties) within the red team should lead this Process.

The attack planner/conductor should derive the details of the options for red teaming based on the information.

In developing attack scenarios, attacks should be constructed based on the configuration of typical defense mechanisms in the AI system.

In addition, the actual reported attack methods, attack trends, actual damage cases, and knowledge of blind spots that are often overlooked in countermeasures should be taken into consideration.

Even if the decision is made to refrain from conducting attack scenarios due to social or other concerns, high risk scenarios should be documented in the report.

5.3.5 Attack deployment

5.3.5.1 Execution procedure

The execution of red teaming should follow a structured, three-step approach:

a) Exploratory Testing (Individual Prompts): The red team shall conduct initial testing using a broad range of attack signatures on individual prompts to identify effective attack methods (e.g., jailbreaking techniques) independent of the system context.

b) Development of Attack Procedures: Based on the results of exploratory testing, the red team shall develop customized attack signatures and payloads tailored to the specific risk scenarios and target system characteristics.

c) System-wide Testing: The red team shall execute the developed procedures against the entire AI system to verify if the malicious output from the AI model successfully compromises downstream components or bypasses post-processing filters.

5.3.6 Impact analysis

5.3.6.1 Severity Assessment Criteria

The red team shall analyze the impact of successful attacks using domain-specific criteria. The analysis must distinguish between "theoretical risk" and "demonstrated impact."

- **CBRN & Physical Safety Impact:**
    - **Actionability:** Does the generated output provide actionable instructions (e.g., a working formula for a pathogen) or merely general information available on Wikipedia?
    - **Novelty:** Does the AI provide novel threat capabilities that lower the barrier to entry for malicious actors?
    - **Severity Level:** Critical (Immediate physical threat), High (Detailed assistance), Low (General knowledge).
- **Performance & Availability Impact:**
    - **Resource Saturation:** Measure the increase in CPU/GPU usage and memory consumption during the attack.
    - **Latency Degradation:** Calculate the response time delay compared to baseline.
    - **Severity Level:** Critical (System crash/DoS), High (>200% latency spike), Medium (Noticeable slowdown), Low (Negligible).
- **Quality & Trustworthiness Impact:**
    - **Hallucination Severity:** Assessing whether the error is benign (minor detail) or dangerous (e.g., incorrect medical dosage, legal misinformation).
    - **Bias & Fairness:** Quantifying the extent of discriminatory output against protected groups.
- **Systemic Impact & Blast Radius:**
    - **Lateral Movement:** Does the compromised agent provide a pathway to access other isolated systems or higher-privilege agents?
    - **Cascading Failures:** Does a failure or hallucination in one agent propagate to other agents or downstream systems, amplifying the impact (e.g., a hallucinated file deletion command executed by a connected tool)?
    - **Traceability Loss:** Is the agent performing actions that cannot be attributed to a specific user or logged event, hindering forensic analysis?

5.3.6.2 Downstream Consequence Analysis

The analysis shall extend beyond the AI model's output to the broader system:

- **Execution Risk:** If the AI output is code (e.g., SQL, Python), was it successfully executed by the backend? (e.g., Did the SQL injection actually retrieve data?)

- **Reputational Damage:** Assessing the potential public relations impact if the failure were public.

5.3.6.3 Root Cause Analysis

The red team should attempt to identify the cause of the failure (e.g., insufficient training data, weak system prompt, lack of output filtering) to inform the remediation plan.

5.3.7 Record keeping during red teaming

The organization should retain red teaming records in accordance with the organization's document management and confidentiality policies.

5.4 Phase 3: knowledge sharing and reporting

5.4.1 Preparing the report of red teaming results

The attack planner shall document the red teaming results in a report. And review it for factual errors, as necessary, with development and provision managers of the target Al system and with other relevant stakeholders.

5.4.2 Test completion report

Development and provision managers of the target Al system should issue the final test completion report based on the red teaming findings.

In an automated environment, comprehensive reports can be generated as a result of red teaming. Automated alerting systems designed to notify stakeholders immediately upon detection of new vulnerabilities are also implemented.

Red teaming scenarios can beoften translated into and saved as test cases. These can be executed regularly (e.g., on a weekly basis) as part of drift testing against the operational model or used during adversarial testing for new models.

5.4.3 Development of improvement plans

The organization shall develop a remediation plan to address identified risks based on the red teaming results.

The plan shall:

— Prioritize remediation measures based on the level of urgency, business risk, and impact.

— Consider a layered defense approach, including pre-filtering inputs, hardening the model (e.g., fine-tuning), and post-filtering outputs.

5.4.4 Broader dissemination

5.4.4.1 Internal Knowledge Transfer (Cross-Team Sharing)

The organization should establish a mechanism to share red teaming findings with other internal AI development teams to prevent the recurrence of similar vulnerabilities.

- **Attack Signature Library:** Create a shared repository of effective attack prompts (e.g., successful jailbreak patterns, sponge attack inputs) that other teams can use for regression testing.

- **Design Patterns for Mitigation:** Share successful defense strategies (e.g., specific system prompt structures that resisted CBRN injection) across the organization.

- **Lesson Learned Sessions:** Conduct workshops to discuss the methodology used to uncover complex logic or quality failures.

5.4.4.2 Executive and Risk Management Reporting

The findings shall be synthesized into a high-level risk report for top management and legal/compliance stakeholders.

- **Residual Risk Profile:** Clearly communicate the remaining risks (e.g., "The model still has a 2% failure rate for sophisticated chemical weapon inquiries") to inform deployment decisions.

- **Compliance Status:** Report on the system's alignment with regulatory requirements (e.g., Safety, Fairness) based on the assessment results.

5.4.4.3 Controlled Dissemination of Sensitive Findings (CBRN/Safety)

For high-risk findings, particularly those related to **CBRN (Chemical, Biological, Radiological, Nuclear)** or **Critical Safety**, the organization shall enforce strict access controls.

- **Need-to-Know Basis:** Detailed attack vectors and generated harmful outputs regarding CBRN shall be restricted to the security team and authorized developers only. They should not be stored in open internal wikis.

- **Sanitized Reporting:** General reports distributed to wider audiences must be sanitized to remove actionable harmful information (e.g., removing the actual chemical synthesis steps generated by the AI).

5.4.4.4 External Responsible Disclosure

If the red teaming identifies vulnerabilities in third-party components (e.g., OSS models, external APIs) or contributes to the broader AI safety community, the organization should follow responsible disclosure procedures.

- **Vulnerability Disclosure:** Adhere to **ISO/IEC 29147** when reporting vulnerabilities to external vendors or coordinating with CERTs.

- **Contribution to Safety Databases:** Consider sharing anonymized attack patterns (not specific vulnerabilities) with industry databases (e.g., AI Incident Database, CVE) to improve global AI safety standards.

5.4.5 Follow-up and re-verification

After implementing improvement measures, the organization shall conduct a follow-up assessment or re-testing to verify that the identified vulnerabilities have been effectively mitigated . Red teaming should be conducted periodically or triggered by significant system updates (e.g., model retraining, architectural changes) to ensure continuous validity of safety measures.

Red teaming should be an iterative process conducted throughout the system lifecycle, rather than a single event prior to deployment.

6 Red teaming techniques

6.1 Overview of AI red teaming techniques

6.1.1 General

A variety of AI red teaming techniques are available to assess and enhance the safety and robustness of AI systems. These techniques range from manual testing by human experts to automated methods leveraging advanced algorithms. Detailed descriptions of these techniques are provided in Annex B.

6.1.2 Perspectives for designing attack scenarios

To effectively identify vulnerabilities in AI systems, particularly those involving Large Language Models (LLMs), attack scenarios should be designed from three distinct perspectives regarding the attack vector and impact:

a) Bypassing pre-processing and input embedding: Focus on techniques to circumvent input filters or embed malicious inputs into reference resources (e.g., indirect prompt injection via RAG sources) to reach the core model.

b) Inducing malicious output: Focus on manipulating the AI model to generate unintended, harmful, or policy-violating outputs (e.g., toxic content, malware code) despite safety training or system prompts.

c) Bypassing post-processing and system impact: Focus on whether the generated malicious output can bypass output filters and execute unintended actions on the broader system (e.g., executing OS commands, SQL injection).

6.1.3 Success Criteria per Dimension

The red team shall define success criteria for attacks based on the target dimension:

- **CBRN/Safety:** Any successful generation of actionable harmful information constitutes a system failure (Zero-tolerance approach).

- **Quality:** The attack is successful if the error rate (e.g., hallucination rate) under adversarial conditions exceeds a defined threshold compared to benign conditions.

- **Performance:** The attack is successful if the input causes latency or resource consumption to spike beyond acceptable operational limits (e.g., >200% increase) compared to standard inputs of similar length.

6.2 Selecting appropriate red teaming techniques

Organizations should select red teaming techniques based on:

— Objectives: Align techniques with specific goals, such as robustness requirements, bias identification or other compliance requirements;

— Resources: Consider the availability of computational resources, expertise, and time;

- System characteristics: Choose methods suitable for the AI model's architecture, data modalities, and functionalities;
- Risk management: Prioritize techniques that address the most critical vulnerabilities identified in the threat modeling phase.

Organizations should refer to Annex B to select appropriate red teaming techniques based on their specific needs, objectives, and resources.

6.3 Selection of testing approach based on system components

The choice between black-box, white-box, or gray-box testing techniques should be determined by the accessibility and ownership of the AI system components:

a) Commercial services (API-based): For components utilizing third-party commercial AI models where internal parameters are inaccessible, black-box testing is primarily applicable. Testing focuses on input-output analysis via APIs.

b) Open-source software (OSS) and in-house development: For components based on OSS or developed in-house, white-box testing is recommended. This allows the red team to utilize internal information such as model weights, gradients, training data, and system prompts to design more effective attacks.

c) Hybrid systems: Complex AI systems often combine both commercial and proprietary components. In such cases, the red teaming plan should segregate components to apply the appropriate testing method (black-box or white-box) for each part .

7 Procedures for documenting and reporting findings

7.1 Data recording and management

7.1.1 Data collection planning

[TBD]

7.1.2 Data structure and storage

[TBD]

7.2 Report structure

7.2.1 Executive summary

[TBD]

NOTE PE note: consider the structured reporting format, covering: - Tested AI system and its scope - Identified vulnerabilities and attack scenarios - Input to the risk prioritization and impact analysis - Mitigation recommendations.

7.2.2 Methodology

[TBD]

7.2.3 Detailed findings

[TBD]

7.2.4 Recommendations

[TBD]

7.2.5 Appendices

[TBD]

7.3 Communication Protocols

7.3.1 Stakeholder identification

[TBD]

7.3.2 Secure communication channels

[TBD]

7.3.3 Feedback and remediation tracking

[TBD]

7.4 Confidentiality maintenance

7.4.1 Data anonymization

[TBD]

7.4.2 Access control policies

[TBD]

7.4.3 Legal compliance

[TBD]

7.5 Differentiation between identification and measurement

[TBD]

8 Integration into AI development and deployment lifecycle

red teaming should be integrated at multiple stages of the AI lifecycle, include:

— Development Phase: Risk assessment before deployment

— Deployment Phase: testing before AI system goes live

— Monitoring Phase: Ongoing adversarial testing and attack simulations.

[TBD]

9 Guidance on adapting traditional software testing methodologies

9.1 Evolution of test case design

9.1.1 Dynamic testing strategies

[TBD]

9.2 Utilization of automated testing tools

9.2.1 AI-specific testing platforms

[TBD]

9.2.2 Scalability considerations

[TBD]

9.2.3 Integration with development tools

[TBD]

9.3 Interpretation of results of red teaming

9.3.1 Statistical significance

[TBD]

NOTE: It cannot be assumed that determining statistical significance alone constitutes a sufficient interpretation of test results

9.3.2 Error categorization

[TBD]

9.3.3 Visualization techniques

[TBD]

Annex A
(informative)

Use case examples

A.1 @@@

PE ACTION-06: propose to modified annex A

Annexes are an optional element of the text.

For rules on the drafting of annexes, refer to the ISO/IEC Directives, Part 2:2021, Clause 20.

To create new annexes, use the Insert Annex function in the ribbon above.

- *Specify whether the annex is normative or informative.*
- *All annexes need to be referred to at least once in the main body of the text.*
- *For an Annex to be considered normative, it needs to be referred to in a requirement in the main body of the text.*

Annex B
(informative)

AI red teaming techniques

B.1 Overview of AI Red Teaming Methods

PE ACTION-07: specific AI red teaming techniques (e.g., model inversion, evasion attacks, prompt injection) 오류! 책갈피 이름이 지정되어 있지 않습니다.

B.2 Typical attack techniques for AI systems

Table B.1 lists typical attack techniques applicable to general AI systems and LLM-based systems.

B.3 Prompt Injection Categories

Prompt injection attacks can be categorized into direct and indirect methods:

a) **Direct Prompt Injection:** The attacker directly inputs malicious prompts to bypass restrictions (e.g., "Ignore previous instructions and do X") . Techniques include role-playing, prefix injection, and special encoding (e.g., Base64).

b) **Indirect Prompt Injection:** The attacker embeds malicious prompts into external data sources (e.g., websites, documents) that the AI system retrieves via RAG or plugins. The AI processes this tainted data, leading to a compromised response without direct attacker interaction with the model .

B.4 Red teaming approaches by domain

[TBD]

Table B.X — Red Teaming Approaches by Domain

Evaluation Domain	**Primary Risk**	**Typical Red Teaming Technique**	**Key Metric**
CBRN / Safety	Creation of hazardous agents, weapons	**Expert Red Teaming:** Domain experts (e.g., biologists) verify if the AI provides novel, actionable paths to harm not easily found via search engines.	Attack Success Rate (ASR), Severity Score
Quality / Accuracy	Hallucination, Unfaithful reasoning	**Knowledge Injection / Counterfactuals:** Injecting false premises or conflicting context (RAG) to test if the model adheres to ground truth.	Factuality Score, Consistency Rate

Evaluation Domain	Primary Risk	Typical Red Teaming Technique	Key Metric
Performance	Availability degradation, High cost	**Sponge Attacks / Complexity Attacks:** Inputs designed to maximize computation time (e.g., highly recursive prompts).	Latency drift (ms), Energy/Cost per query
Agent Autonomy	**Goal Manipulation & Loop Evasion**	**Instruction Injection via Environment:** Injecting conflicting goals into the environment (e.g., a file named "DO_NOT_DELETE.txt" containing instructions to delete all files) to test goal adherence.	Task Completion Rate, Deviation Rate
Resource & Cost	**Economic DoS / Exhaustion**	**Infinite Loop Induction:** Crafting inputs that force the agent into a recursive reasoning loop or excessive tool usage to exhaust API quotas or compute budgets.	Resource Consumption, Cost per Task
Multi-Agent	**Trust Exploitation**	**Spoofing & Masquerading:** Impersonating a trusted agent or orchestrator to issue unauthorized commands in a multi-agent system.	Authentication Bypass Rate

Annex C
(informative)

Examples of document templates

C.1 Test plan

C.2 Recommendation report/remediation plan

C.3 Higher-level lessons learned/observations report

C.4 Communication plan

Annex D
(informative)

Example factors that can cause previously unknown vulnerabilities to emerge

D.1 General

It is essential to continuously update and maintain the repository of red teaming cases as AI applications evolve. Real-world interactions expose the AI system to new types of data and scenarios, potentially uncovering previously unknown vulnerabilities.

D.2 Example factors

Examples of factors that can cause previously unknown vulnerabilities to emerge include: - Company content changes: Updates or modifications to the Retrieval-Augmented Generation (RAG) knowledge base or alterations in the company’s products.

- Evolving news and events: Real-world events occurring after the foundational model's training phase, such as major global events (e.g., the 2024 Olympic Games, new CEO appointments, or significant political events such as U.S. elections).

- Advancements in cybersecurity research: Identification of new prompt injections and vulnerabilities discovered through ongoing research by the cybersecurity community.

- Introduction of new AI model versions: Changes in model architecture, prompts, foundational model updates, or alterations in the model’s behavior.

Regular, proactive updates to red teaming scenarios and comprehensive reporting of findings ensure ongoing AI system resilience, adaptability, and compliance with regulatory and ethical standards.

Annex E
(informative)

Relationship with ISO/IEC/IEEE 29119-2

E.1 General

This annex describes the alignment between the AI red teaming processes defined in this document and the software testing processes defined in ISO/IEC/IEEE 29119-2. The AI red teaming methodology follows a specific three-phase approach (Team formation and preparation, Execution, and Knowledge sharing), which maps to the Test Management Processes and Dynamic Test Processes of ISO/IEC/IEEE 29119-2.

E.2 Process Mapping

Table E.1 provides a mapping between the phases and activities of ISO/IEC AWI TS 42119-7 and the corresponding processes in ISO/IEC/IEEE 29119-2.

Table E.1 — Alignment between ISO/IEC 42119-7 and ISO/IEC/IEEE 29119-2

ISO/IEC 42119-7 Phase and Activities	ISO/IEC/IEEE 29119-2 Process	Mapping Explanation
Phase 1: Team formation and preparation • Team selection • Objective setting • Scope definition • Development of risk scenarios	**7.2 Test strategy and planning process**	The preparation phase corresponds to the test planning activities. Red team formation aligns with staffing (TP6). Identifying risk scenarios aligns with risk identification and analysis (TP3). Scope definition and objective setting are core parts of the test plan development.
Phase 1: Target Environment Preparation • Preparing the environment • Adjusting monitoring configurations	**8.3 Test environment and data management process**	The red team must establish a test environment (e.g., staging vs. production) and prepare necessary access (API keys, logs). This maps directly to establishing the test environment (ED1) in 29119-2.
Phase 2: Execution of AI red team activities • Step 1: Exploratory Testing (Individual Prompts)	**8.2 Test design and implementation process**	Exploratory testing in red teaming involves identifying effective attack vectors. This aligns with deriving test cases (TD3) based on a test model (TD1), although in exploratory testing, design and execution are often concurrent.
Phase 2: Execution of AI red team activities • Step 2: Developing attack signatures and procedures	**8.2 Test design and implementation process**	Developing specific "attack payloads" and detailed procedures based on initial findings maps to creating test procedures (TD4) in 29119-2.

Phase 2: Execution of AI red team activities • Step 3: System-wide Testing	**8.4 Test execution process**	Executing the developed attack procedures against the full AI system and observing the results maps to executing test procedures (TE1) and comparing test results (TE2).
Phase 2: Escalation and Emergency Procedures • Stop/Go Criteria • Monitoring operational risks	**7.3 Test monitoring and control process**	Monitoring the red teaming exercise for critical failures or operational risks and deciding whether to suspend testing aligns with the control activities (TMC3) and monitoring activities (TMC2).
Phase 2: Incident Reporting • Reporting critical vulnerabilities	**8.5 Test incident reporting process**	Immediate reporting of critical vulnerabilities found during execution maps to creating incident reports (IR2) for further action.
Phase 3: Knowledge sharing and reporting • Preparing the report • Development of Improvement Plans	**7.4 Test completion process**	The final phase of creating a report, sharing findings, and identifying lessons learned maps directly to the test completion reporting (TC4) and identifying lessons learned (TC3).

Bibliography

[1] ISO/IEC TS 42119-2:2025, *Artificial intelligence — Testing of AI — Part 2: Overview of testing AI systems*

[2] ISO/IEC 5338:2023, *Information technology — Artificial intelligence — AI system life cycle processes*

[3] ISO/IEC/IEEE 16085:2021, *Systems and software engineering — Life cycle processes — Risk management*

[4] ISO/IEC 20246:2017, *Software and systems engineering — Work product reviews*

[5] ISO/IEC TS 25058:2024, *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guidance for quality evaluation of artificial intelligence (AI) systems*

[6] ISO/IEC DIS 25059, *Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality models for AI systems*

[7] ISO/IEC 29147:2018, *Information technology — Security techniques — Vulnerability disclosure*

[8] OWASP GenAI Red Teaming Guide, https://genai.owasp.org/resource/genai-red-teaming-guide/

[9] CSA, Agentic AI Red Teaming Guide, https://cloudsecurityalliance.org/artifacts/agentic-ai-red-teaming-guide