

Glossary

Edition 202608

Copyright © EXIN Holding B.V. 2026. All rights reserved.
EXIN® is a registered trademark of EXIN.

No part of this publication may be reproduced, stored, utilized or transmitted in any form or by any means, electronic, mechanical, or otherwise, without the prior written permission from EXIN.

Introduction

This is the glossary for the EXIN AI Security Professional based on the OWASP AI Exchange certification.

It provides definitions of the basic concepts listed in the preparation guide and used in exam, to support candidates in their exam preparation.

The definitions provided are aligned with the OWASP AI Exchange.

Glossary

The terms below are listed alphabetically, together with their definitions.

Term	Definition
accountability	The principle that individuals or organizations are answerable for their actions and decisions, and can be required to explain, justify, or take responsibility for outcomes.
adversarial attack	Overloaded term, sometimes used to refer to any attack, and sometimes to an attempt to deliberately craft or manipulate input, data, or a model in order to make a machine learning system produce incorrect, harmful, or attacker-chosen output.
adversarial testing	Testing by performing adversarial attacks, in order to uncover exploitable weaknesses before real attackers do.
agentic AI	Systems built around AI agents that can plan, take actions, and interact with tools, data, or other agents to accomplish goals with limited human intervention.
AI (artificial intelligence)	The capability of a computer system to perform tasks that typically require human-like intelligence, such as reasoning, learning, pattern recognition, or creating content, whether through rule-based methods or machine learning.
AI agent	A software component that can autonomously decide on and execute actions, such as calling functions or tools, to achieve a goal, rather than only producing a response.
AI assets	The AI-specific elements belonging to a particular system or organization that need protecting, such as training data and model parameters.
AI jailbreak	A prompt-injection attack bypassing a model's built-in safety or content restrictions to make it produce output it was designed to refuse.
AI management system (AIMS)	An organizational framework of policies and processes, as for example defined by ISO/IEC 42001, for governing AI responsibly and securely.
AI model	A mathematical or computational artifact that maps inputs to outputs, learned from data or defined by rules.
AI model provider	An organization or party that develops, trains, or supplies an AI model for others to use or build upon.

Term	Definition
AI privacy principles	A set of principles for how personal data should be handled, covering aspects such as accuracy, consent, minimization, explainability, fairness, transparency, and limitation.
• **accuracy**	In the context of personal data: the principle that personal data should be accurate and, where necessary, kept up to date. In the context of models: degree to which an AI system's outputs are correct.
• **consent**	A data subject's freely given, specific, informed, and unambiguous agreement to the processing of their personal data.
• **data minimization and storage limitation**	The principle of collecting only the data necessary for a specific purpose and retaining it no longer than needed for that purpose.
• **data subject rights**	Individual rights over personal data, such as access, correction, and deletion.
• **explainability**	The principle that individuals should be informed about how their data is used and, where automated decisions affect them, be able to understand through wat steps those decisions were reached.
• **fairness and lawfulness**	The principle that personal data must be processed in ways that are legally justified and do not unjustly disadvantage individuals or groups.
• **transparency**	The principle that individuals should be informed about how their data is collected, used, shared, retained, and, where automated decisions affect them, be able to understand the approach for reaching those decisions.
• **use limitation and purpose specification**	The privacy principle that personal data collected for one purpose should not simply be reused to train an AI model for a different, incompatible purpose without a proper lawful basis.
AI red teaming	A structured, adversarial testing exercise in which testers act like real attackers to uncover security, safety, or ethical weaknesses in a system.
AI security	The discipline of protecting AI systems, their data, and their supporting infrastructure against security threats that exploit characteristics specific to AI, in addition to conventional cybersecurity risks.
AI security governance	The organizational structures, policies, and oversight processes that ensure AI-related security risks are identified, managed, and accounted for.
AI security testing approach	A structured methodology for planning and executing security tests against an AI system, covering scope, techniques, and success criteria.
AI system	A complete system built around one or more AI models, including its data pipelines, interfaces, and supporting infrastructure.

Term	Definition
AI system provider	The organization that develops or assembles an AI system and makes it available for deployment or use.
anonymization	Irreversibly processing personal data so that the individual it relates to can no longer be identified, even indirectly.
API (application programming interface)	A defined set of protocols and tools that allows different software programs or components to communicate and exchange functionality.
audit	An independent review of a system, process, or organization to verify compliance with standards, policies, or regulations.
augmentation data	External data added to a model's context at some point in its lifecycle to improve or extend its output, such as documents retrieved for retrieval-augmented generation.
automated oversight	Monitoring or control of a system carried out by automated tools rather than by people.
automation	The use of technology to perform tasks or processes with reduced human intervention.
bias	A systematic skew in data, decisions, or outputs often introduced unintentionally through data or design choices. Depending on the case, bias may lead to unfair or inaccurate treatment of certain groups or cases.
blast radius	The extent of damage, compromise, or exposure that would result if a particular asset, component, or credential were successfully attacked or failed.
chatbot	A software application that interacts with users through text or voice conversation, automating responses that would otherwise require a human.
Chief AI Officer (CAIO)	The organizational role responsible for an organization's overall AI strategy, governance, and risk oversight.
Chief Information Security Officer (CISO)	The senior executive responsible for an organization's overall information and cybersecurity strategy and program.
code	The set of instructions, written in a programming language, that implements the logic of a software program.
compliance	Adherence to applicable laws, regulations, standards, and internal policies governing an organization's activities.
confidentiality	The property that information is accessible only to those authorized to access it. One of the three core objectives of information security alongside integrity and availability.
configuration leak	The unintended exposure of a system's configuration details, such as settings or credentials, to unauthorized parties.

Term	Definition
copyright	A form of legal protection granted to the creators of original works, giving them exclusive rights over the reproduction, distribution, and use of those works.
data leak	The unauthorized or unintended exposure of data to parties who should not have access to it.
data protection	In the context of personal data, measures and practices intended to safeguard personal data from unauthorized access, use, or disclosure.
data protection impact assessment (DPIA)	A structured process, required under data protection law, for identifying and mitigating privacy risks before carrying out a type of data processing likely to result in high risk to individuals.
data security	Protecting data against unauthorized access, alteration, or loss throughout its lifecycle.
data subject	An identified or identifiable living individual to whom personal data relates.
denial-of-service (DoS)	An attack that aims to make a system or service unavailable to legitimate users, typically by overwhelming it with traffic or requests.
denial-of-wallet (DoW)	An attack that deliberately drives up a target's operating costs, for example by triggering excessive resource-intensive requests against a pay-per-use service, causing financial rather than availability harm.
deployer	A person or organization that puts a system into operational use, as distinct from the party that developed it.
development-time	The phase in a system's lifecycle when it is designed, built, and trained or tested, as opposed to when it is deployed and running.
• **development-time security threats**	Security risks that can be introduced while a system is being designed, built, or trained, such as compromised code, backdoored components, or corrupted data.
• **data poisoning development-time**	Deliberately corrupting or manipulating the data used to train a machine learning model so that the resulting model learns flawed or attacker-controlled behavior.
• **model poisoning development-time**	Tampering directly with a model's parameters, architecture, or training process to embed hidden or malicious behavior.
• **sensitive data leak development-time**	The unintended exposure of sensitive data, such as source data, credentials, or configuration secrets, during a system's development process.
embedding	A numerical vector representation of data, such as text or images, that captures its semantic meaning so that machine learning models can process, compare, or search it.

Term	Definition
encryption	The process of encoding data so that it can only be read by parties who hold the correct decryption key, used to protect data confidentiality.
ethical principles	A set of values, such as fairness, honesty, and accountability, that guide what is considered right or acceptable conduct in a particular field or activity.
European Union (EU)	A political and economic union of European member states that adopts laws and regulations, such as the GDPR and the AI Act, binding on its members.
G.U.A.R.D (Govern, Understand, Adapt, Reduce, Demonstrate)	The five-step framework by the OWASP AI Exchange (Govern, Understand, Adapt, Reduce, Demonstrate) for structuring an organization's approach to managing AI security risk.
General Data Protection Regulation (GDPR)	The European Union regulation that sets rules for the protection and processing of the personal data of individuals.
generative AI	AI systems that create new content, such as text, images, or code, typically by learning patterns from large training datasets.
hijacking	An attack in which an adversary takes unauthorized control of a system, session, resource, or communication channel and redirects it toward their own purposes.
human oversight	The involvement of people in monitoring, reviewing, or intervening in the decisions and actions of an automated or autonomous system.
inference-time	The stage in a machine learning model's lifecycle at which it is actively used to produce outputs from new inputs, as opposed to training.
information security management system (ISMS)	A systematic framework of policies, processes, and controls, as defined in ISO/IEC 27001, for managing an organization's information security risks.
input data	Data supplied to a system or process in order to produce an output. In the context of AI, the data that is used as input for the AI model.

Term	Definition
input security threats	Security risks that arise from data or content supplied to a system as input, such as malicious or manipulated content designed to alter its behavior.
• **evasion**	An attack that crafts malicious input in terms of data, as opposed to instructions, specifically designed to bypass a detection or classification system, causing it to produce an incorrect result. For example, a subtly altered traffic sign that a vision model misreads.
• **model exfiltration**	Extracting a copy or functional equivalent of a proprietary model, often through repeated queries, without authorization.
• **prompt injection**	An attack that manipulates the instructions given to a generative AI model, through the prompt or injected content, to override intended behavior.
• **resource exhaustion**	An attack or condition that consumes computational, memory, or storage, to the point of degrading or denying service to legitimate users or causing large financial impact.
• **sensitive data disclosure through use**	Disclosing of sensitive training or other data through normal or abusive use of the system.
integrity verification	Checking that data, code, or another digital artifact has not been altered or tampered with, typically using techniques such as checksums, hashes, or digital signatures.
intellectual property (IP)	Legally protected creations of the mind, such as trade secrets, patents, copyrighted works, and trademarks, over which their creator or owner holds exclusive rights.
ISO/IEC 42001	An international standard specifying requirements for an AI management system to govern AI responsibly and securely.
large language model (LLM)	A machine learning model trained on vast amounts of text that can interpret and generate natural language.
leakage	The unauthorized or unintended disclosure of information that should have remained confidential or protected.
machine learning	A branch of AI in which systems learn patterns from data rather than being explicitly programmed with rules.
membership inference	An attack that determines whether a specific record was part of a model's training data, potentially revealing sensitive information about individuals.
model architecture	The structural design of a model, including its components and how they are connected.
model inversion	An attack that reconstructs or approximates sensitive training data by analyzing a model's outputs.

Term	Definition
model leak	The unauthorized exposure or theft of a model's parameters, weights, or architecture.
open source	Software or other technical artifacts whose underlying source code or components are made publicly available for anyone to use, inspect, modify, and redistribute.
output data	The result produced by a system or process in response to a given input. In the context of AI, typically the output of a model.
output sanitization	Filtering or cleaning a system's output before it is passed on or displayed, to remove harmful, sensitive, or malicious content.
output validation	Checking a system's output against defined criteria to ensure it is correct, safe, and fit for its intended use before it is acted upon or passed downstream.
personal data	Any information relating to an identified or identifiable living individual.
predictive AI	AI systems designed to forecast outcomes or classify data based on patterns learned from historical data, as opposed to generating new content.
privacy	An individual's right to control how their personal information is collected, used, and shared.
privacy by default	The principle that a system's default settings should be the most privacy-protective option available, requiring no extra action from the individual, a companion concept to privacy by design.
privacy by design	Building privacy protections into a system's architecture and processes from the earliest stages of development, rather than adding them afterward.
privacy impact assessments (PIA)	An assessment used to identify and address the privacy risks that a project, system, or process may pose to individuals.
privacy incident	An event in which personal data is processed, accessed, or disclosed in a way that violates privacy protections.
processing (of personal data)	Any operation performed on personal data, such as collection, storage, use, or deletion.
prompt	The input text or instructions given to a generative AI model to elicit a response.
pseudonymization	Replacing identifying information in data with artificial identifiers, so that additional information, which is held separately, is needed to re-identify the individual.
publicly accessible	Data, systems, or resources that can be accessed by anyone without authentication or special permission.

Term	Definition
ready-made model	A pretrained model obtained from a third party and used as-is or with adaptation, rather than trained from scratch.
residual risk	The risk that remains after security controls and mitigations have been applied.
responsible AI	An approach to developing and operating AI systems that accounts for ethics, fairness, safety, and security alongside performance.
retrieval-augmented generation (RAG)	A technique in which a generative AI model retrieves relevant external documents or data at inference time to ground and enrich its responses.
risk	The potential for a threat to exploit a vulnerability and cause harm, typically assessed in terms of likelihood and impact.
risk analysis	The process of identifying and evaluating risks to determine their likelihood and potential impact.
risk levels for AI systems based on AI Act	The EU AI Act's classification of AI systems into tiers of regulatory risk, which determines which legal obligations apply to a given system.
• **minimal or no risk**	The AI Act tier for AI systems considered to pose negligible risk to health, safety, or fundamental rights, subject to little or no regulation.
• **limited risk**	The AI Act tier for AI systems that pose some risk, typically requiring specific transparency obligations, such as disclosing that content is AI-generated.
• **high risk**	The AI Act tier for AI systems that can significantly affect health, safety, or fundamental rights, subject to strict obligations before and after deployment.
• **unacceptable risk**	The AI Act tier for AI practices considered to pose such severe risk that they are prohibited outright.
risk management	The ongoing process of identifying, assessing, treating, and monitoring risks to keep them within acceptable levels.
risk mitigation	The selection and implementation of measures to reduce the likelihood or impact of an identified risk.
risk rating	A score or classification assigned to a risk based on its assessed likelihood and impact.
runtime	The phase in a system's lifecycle when it is deployed and actively operating, processing real inputs in a live environment.

Term	Definition
runtime security threats	Security risks that arise while a system is operating in a live, production environment, as opposed to during development.
• **augmentation data manipulation**	Tampering with the external data a system retrieves or injects at runtime, such as retrieval-augmented generation sources, to influence its output.
• **direct augmentation data leak**	The exposure of augmentation data through unauthorized direct access in transit or at rest.
• **direct runtime model leak**	The unauthorized exposure of a model's parameters or architecture through unauthorized direct access in transit or at rest.
• **direct runtime model poisoning**	Attacker manipulation of a deployed model's behavior through unauthorized direct access at runtime, in transit or at rest.
• **input data leak**	The unauthorized exposure of data supplied as input to a system.
• **output containing conventional injection**	Output from a system that itself contains malicious content, such as script or query fragments, which triggers a conventional injection attack, such as SQL injection or cross-site scripting, when passed to a downstream system.
safety of an AI system	The property that an AI system does not cause unacceptable harm to people, property, or the environment through its operation.
security breach	An incident in which security controls are bypassed or fail, resulting in unauthorized access, disclosure, or damage.
security controls	The technical, procedural, or organizational measures put in place to reduce security risk.
• **controls for sensitive data limitation**	Security controls specifically aimed at restricting the collection, exposure, or retention of sensitive data.
• **controls to limit the effect of unwanted behavior**	Security controls designed to contain or reduce the impact of a system's unintended or harmful behavior once it occurs.
• **governance controls**	Organizational-level controls, such as policies, oversight structures, and accountability mechanisms, that manage risk at a strategic level.
(security) attack	A deliberate action by an adversary intended to compromise the confidentiality, integrity, or availability of a system or its assets.
security incident	An event that indicates a possible breach of security policy, control failure, or compromise of an asset.
security risks	The potential for security threats to exploit vulnerabilities and cause harm.

Term	Definition
security testing	The process of evaluating a system to identify security weaknesses before they can be exploited.
security threats	The full catalog of ways a system's confidentiality, integrity, or availability can be attacked. In context of AI systems, this includes spanning input threats, development-time threats, and runtime conventional threats.
sensitive data	Data that requires special protection due to its potential for harm if exposed, such as personal, confidential, or classified information.
service provider	A third party that supplies a service, such as hosting, infrastructure, or software, on which an organization or system depends.
source code	The human-readable instructions, written in a programming language, that make up a software program.
source code leak	The unauthorized exposure of a system's source code to unauthorized parties.
sponge attack	An attack that feeds a model inputs designed to maximize its computation time or resource use, degrading its performance or availability.
stakeholder	Any individual or organization with an interest in, or affected by, a project, system, or decision.
startup	An early-stage company, typically founded to develop and scale a new product, technology, or business model.
structured query language (SQL)	A standard language used to create, query, and manage data held in relational databases.
tampering	Unauthorized modification of data, code, or a system to alter its behavior or integrity.
third party	Any external organization or individual, other than the primary provider and user, that is nonetheless involved in or affects a system, such as through its supply chain.
threat modeling	A structured process for identifying potential threats, vulnerabilities, and attack paths against a system, used to prioritize security measures.
token	The basic unit of text, such as a word, part of a word, or character, into which language is broken down for processing by a language model. In cloud AI service contexts, tokens are also the billing unit.
tokenization	The process of splitting text into smaller units for processing by a language model, or of replacing sensitive data elements with non-sensitive placeholder values to protect the original data.
tool	In the context of agentic AI, an external function, API (application programming interface), or capability that an AI agent can invoke to perform actions beyond generating text.

Term	Definition
training data	The data used to train a machine learning model to recognize patterns or perform its task.
trustworthy AI	AI that is designed and operated to be lawful, ethical, and technically robust, including being secure and safe.
unwanted behavior	Any output, decision, or action from a system that deviates from its intended, safe, or expected behavior.
vulnerability	A weakness in a system, process, or control that could be exploited by a threat to cause harm.

Certified for what's next

Contact EXIN

www.exin.com