RÉPUBLIQUE
FRANÇAISE
Liberté
Égalité
Fraternité

SECURITY RECOMMENDATIONS FOR A GENERATIVE AI SYSTEM

ANSSI GUIDELINES

ANSSI-PA-102
29/04/2024

TARGETED AUDIENCE:

Developers **Administrators** **IT security managers** **IT managers** **Users**

Information

Warning

This document, written by ANSSI, the French National Cybersecurity Agency, is titled **"Security recommendations for a generative AI system"**. It is freely available at `cyber.gouv.fr/en`.

It is an original creation from ANSSI and it is placed under the "Open Licence v2.0" published by the Etalab mission.

According to the Open Licence v2.0, this document can be freely reused, subject to mentionning its paternity (source and date of last update). Reuse means the right to communicate, distribute, redistribute, publish, transmit, reproduce, copy, adapt, modify, extract, transform and use, including for commercial purposes

The recommendations are provided as is and are related to threats known at the publication time. Considering the information systems diversity, ANSSI cannot guarantee direct application of these recommendations on targeted information systems. Applying the following recommendations shall be, at first, validated by IT administrators and/or IT security managers.

This document is a courtesy translation of the initial French document **"Recommandations de sécurité pour un système d'IA générative [29]"**, available at `cyber.gouv.fr`. In case of conflicts between these two documents, the latter is considered as the only reference.

Document changelog:

VERSION	DATE	CHANGELOG
1.0	29/04/2024	First version

Contents

1
Context

1.1 Introduction

Artificial intelligence (AI) has long been a subject of research, but the potential offered by compute resources and big data has opened up new opportunities. These include a significant increase in the number of products that can generate a response to a question in natural language using a model trained on very large volume of data. These AI models are generally referred to as Large Language Models (LLMs) and fall into the category of generative AI (see definitions in section 1.2).

The recent enthusiasm for these products and services, some of which have been made easily accessible to the public, has prompted organisations (businesses, government departments) to look at the potential productivity benefit that could be achieved using AI.

While this technology offers new opportunities for the organisation of work, it is important to be vigilant and cautious in the approach to the deployment and integration of AI in an existing information system. The deployment of generative AI tools gives rise to new threats that could have a significant impact, for example on the confidentiality of the data processed by these tools, but also on the integrity of the information systems to which they are connected.

The purpose of this document is to provide security recommendations for the use of generative AI solutions based on LLMs within public and private entities.

1.2 Definitions

Generative AI

Generative AI is a subsection of artificial intelligence, focused on creating models trained to generate content (text, images, videos, etc.) from a specific corpus of training data.

Large Language Model

A category of generative AI models that can generate text close to natural human language and which are generally trained on a large dataset.

AI Model

In the context of this guide, an AI model refers to a neural network and its parameters (weights, bias [1]).

AI System

An AI system encompasses all the technical components of an application based on an AI model: implementation of this AI model, front-end services for users, databases, logging, etc.

Query

A query (or prompt) refers to the instruction in text form sent by the user to the AI system.

Adversarial Attack

An adversarial attack aims to send to an AI system one or more malicious requests with the aim of misleading or altering its proper operation.

1. In a neural network, a weight is a power coefficient of the connection between 2 neurons, which is adjusted throughout the training phase. A bias is a constant linked to a neuron that allows for "compensation" in the calculation of the result.

1.3 Scope

This document deals mainly with the following use cases:

- Producing a summary of documentation;
- Retrieval of information or generating text from a corpus of documents;
- Chatbot [2];
- Source code generation for software developers.

The identified documentary corpus could be multi-modal, i.e. it may involve multiple categories of input data: text, image, sound, video, etc. However, the guide focuses primarily on the generation of output text and does not deal specifically with image or video generation (although most of the recommendations also apply to these use cases).

This documentary corpus includes the model's training data, but can also include additional data or documents supplied directly as input by the user.

This document only deals with securing the architecture of a generative AI system based on an LLM.

Security issues related to data *quality* [3] and *performance* [4] of an AI model are not covered in this document.

Similarly, if other issues such as ethics, privacy, intellectual property, the protection of business secrecy, or the protection of personal data are also issues to be taken into account when designing an AI model, these do not fall within ANSSI's area of expertise and are therefore not covered in this guide.

For all of these subjects, you can consult the work of the following organisations: ENISA [6, 7], BSI [1], NIST [15, 16] or CNIL [2].

ANSSI also co-signed an NCSC-UK [14] paper on AI security in November 2023.

2. A chatbot is defined here as an application enabling a written exchange between the user and the AI system rather than an oral exchange.

3. Data quality generally refers to a business criterion. Data quality criteria from a business point of view could be, for example its origin, quantity, completeness, relevance, accuracy, representativeness (in a statistical sense), or conformity to a given structure.

4. The performance of an AI model is also a business concept that is highly dependent on the objectives set when the model was designed. It can include a number of factors such as the accuracy, relevance or speed of responses generated for users, for example.

2
Summary

The implementation of a generative AI system can be divided into 3 cyclical phases: an initial training phase for the AI model based on specifically selected data, then an integration and deployment phase, and finally an operational production phase in which users can access the trained AI model via the AI system.

Each of these 3 phases must be covered by specific security measures, which depend in part on the outsourcing decision for each component (hosting, model training, performance testing, etc.) as well as the sensitivity of the data used in each phase and the criticity of the AI system regarding its business purpose.

In addition to the traditional threats inherent in any information system, an AI system may be subject to specific attacks, for example disrupting its proper functioning (adversarial attacks) or to exfiltrate data processed by itself.

The issue of data security, particularly training data, is therefore a key challenge for a generative AI system, linked with the information accessed by users when they query the model. Actually, it is designed to generate an answer from all the data accessed during training, as well as additional data that may come from sensitive internal sources.

The use of a generative AI system must therefore meet confidentiality requirements (*sending sensitive data to public tools available on the Internet* [5] *must be unauthorized.*) but also meet integrity and availability requirements. AI system interactions with other applications or IS components must therefore be secured, limited to strictly operational needs, and human validation must be implemented when it's critical to the organisation.

Specific uses cases, such as AI-assisted application development, raise a number of major issues and must therefore be managed (with great vigilance over sensitive modules or applications), checked by humans and tested regularly (with automatic source code analysis tools).

Finally, protecting AI models themselves may be as much of an issue as data security not only for reasons of protecting scientific and technical capabilities (academic research, models used for defense and national security, etc.), but also because an attacker with knowledge of the architecture and parameters of an AI model may be able to improve its attack capabilities for other purposes (data exfiltration, etc.).

5. Such as *ChatGPT*, *Gemini* or even *DeepL* for translation.

3
Description of a generative AI system

Warning

The lifecycle and architecture presented in this chapter are given as examples to make the recommendations easier to understand. They are therefore not intended to be prescriptive. In particular, the sequence of the functions presented here is only one possible option. Depending on the use case, these functions may not always be implemented in an AI system.

3.1 Lifecycle of a generative AI system

Figure 1 describes an example of the lifecycle of a generative AI system.

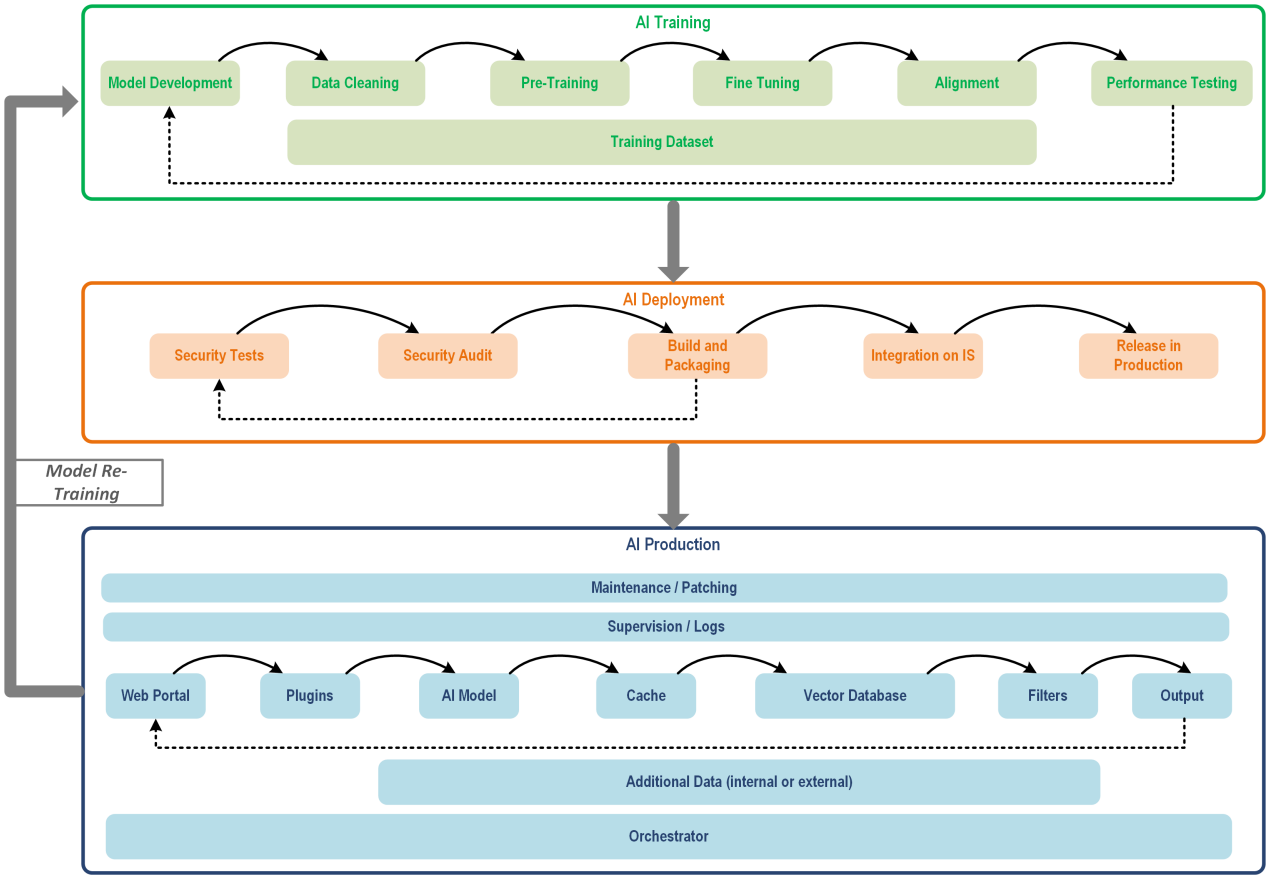

Figure 1 – Example of the lifecycle of a generative AI system

The 3 phases of training, deployment and production [6] may involve different technical environments and different users. It is important that security is treated in each of these 3 phases of a generative AI system's lifecycle.

These 3 phases can be carried out in different environments, for example the training phase in a public Cloud and the deployment and production phases locally within the entity. Nevertheless, appropriate security measures must be applied regardless of the environment chosen.

Re-training an AI model structured in this way does not generally involve repeating all the steps in the training phase (very often, only the fine-tuning or alignment stages are carried out).

Figure 2 shows examples of responsibility sharing throughout the design phases of a generative AI system.

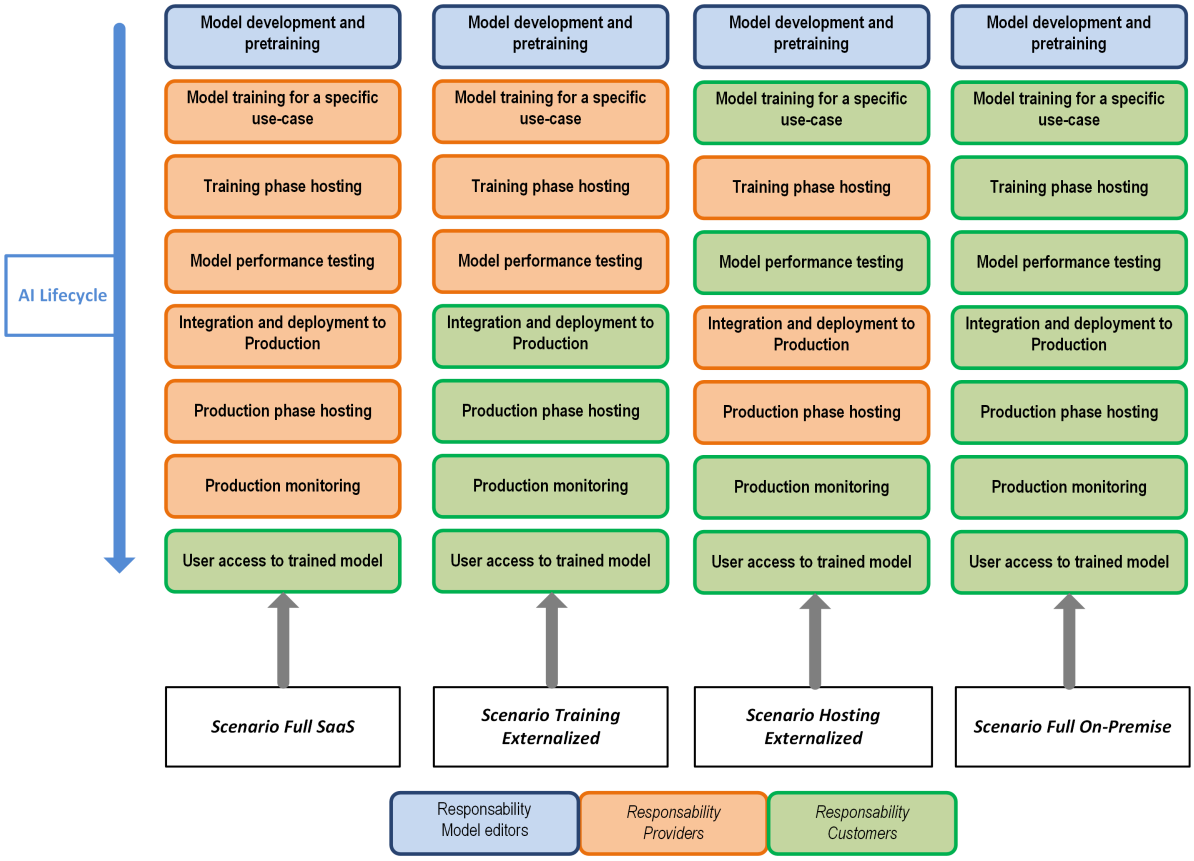

Figure 2 – Shared responsibility scenarios for a generative AI system

The security risks and impacts should be evaluated according to the scenario chosen by the organisation.

6. The production phase may sometimes be called inference phase of the AI model, i.e. the model performs predictions for given users.

Figure 3 describes the integration of a generative AI system into an IS and the important points regarding the internal and external interactions.

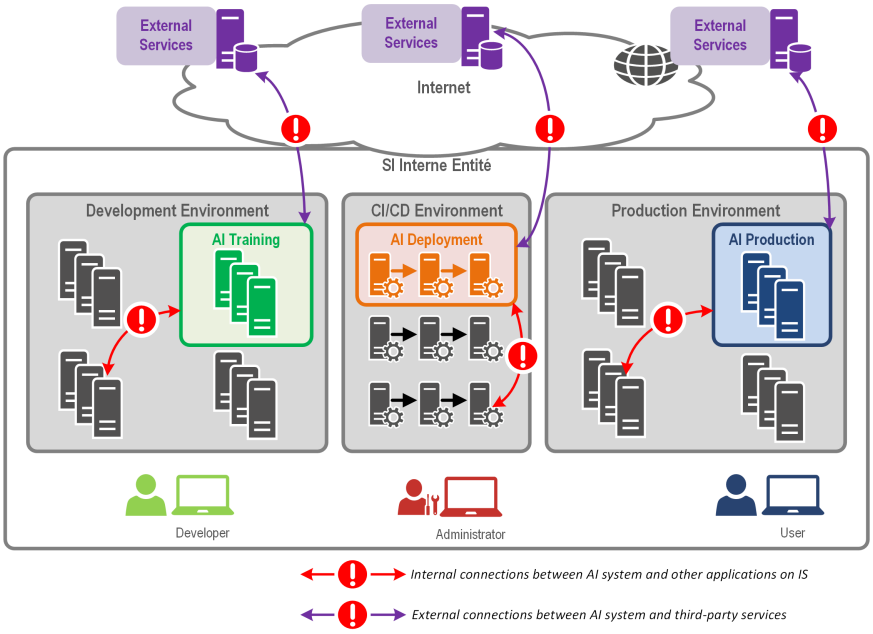

Figure 3 – Integration of a generative AI system into an existing IS

Particular attention must be paid to these interactions and these must be included in the scope of the analysis in all phases of the project.

R1

Integrate security into all phases of the lifecycle of an AI system

Security measures must be identified and applied in each of the 3 phases of the AI system's lifecycle: training, deployment and production. These measures depend heavily on the responsibility-sharing framework adopted and the associated subcontracting. They must also take into account interactions with other applications or components both within and external to the IS.
Refer to the ANSSI cybersecurity hygiene guide [17] to ensure the basic level of security which is to be applied.

3.2 Architecture of a generative AI system

Figure 4 describes an example of the general architecture of a generative AI system.

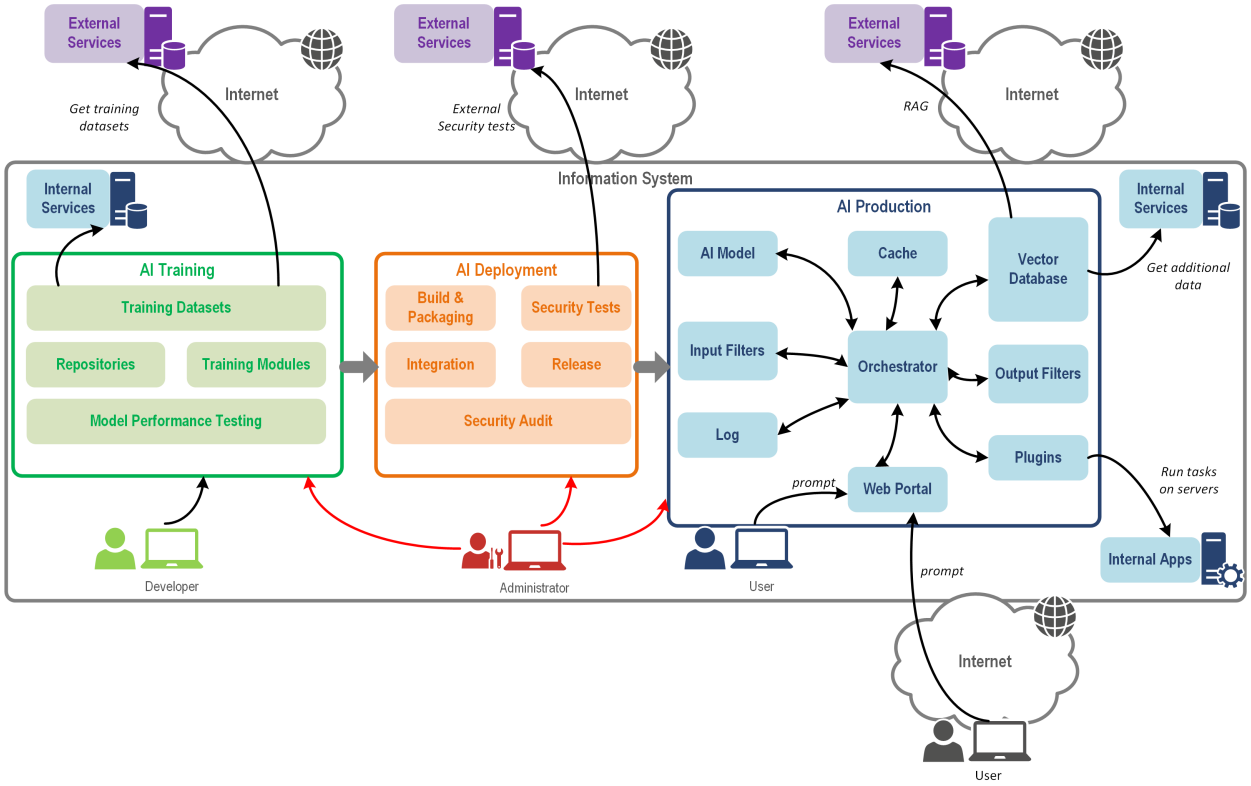

Figure 4 – General architecture of a generative AI system

This architecture is not an exhaustive list of all the components of a generative AI system, but is designed to identify potential attack paths used to target an entity.

There are a number of important elements in this schema:

- the different populations with access to an AI system in each phase: users, developers, administrators, auditors, etc.;
- the vector database, which is generally used to store additional data index in the form of vectors, in order to enrich [7] user queries before sending them to the model (concept known as RAG - Retrieval Augmentation Generation). This database can be built from internal data sources of the organisation or external from partner sources;
- the input and output filters of the AI model, providing defence-in-depth against malicious requests or undesired behaviour by the AI system;
- the plugins or additional components, which can be used to connect the AI system to other business or technical resources within or outside the organisation.

7. A vector database is used in particular in the context of LLMs, because it can be used to establish comparisons and identify relationships between objects, and therefore understand better the context.

4
Attack scenarios on generative AI

A generative AI system is at first a standard business application, which must have the same security baseline as any other business application within the entity. However, in addition to this security baseline, the entity must take into account specific threats to a generative AI system.

These threats can be divided into 3 main categories of attacks [8]:

- **Manipulation attacks**: these attacks involve hijacking the behaviour of an AI system in production by malicious queries. These can lead to unexpected responses, dangerous actions or a denial of service;
- **Infection attacks**: these attacks involve infecting an AI system during its training phase, by altering the training data or inserting a backdoor;
- **Exfiltration attacks**: these attacks involve stealing information about an AI system in production, such as the data used to train the model, user data, or internal model data (parameters).

In the context of generative AI, these attacks can affect the following security requirements:

- **Confidentiality**: the aim is to protect an AI system against the sensitive data leaks: training datasets, user queries, model parameters, additional internal data, etc.;
- **Integrity**: the objective is to protect an AI system against an unexpected change in its behaviour. Integrity can concern the model itself (parameters) or target the training datasets (poisoning) or even the technical components that enable the AI system to work properly: scripts [9], external libraries (supply-chain attack), services configurations, etc. ;
- **Availability**: the objective is to protect an AI system against denial of service or actions intended to degrade its performance (malicious requests);
- **Traceability**: the objective is to guarantee explicability [10] and the accountability of actions carried out on an AI system. These elements can facilitate the investigation and remediation work following a security incident.

8. These categories are from the CNIL taxonomy on this subject: https://linc.cnil.fr/petite-taxonomie-des-attaques-des-systemes-dia.

9. These scripts could be, for example, scripts for *fine-tuning* the AI model or scripts for the deployment or maintenance of an AI system.

10. As defined by CNIL, explicability is the ability to make links between the objects and identify the criteria taken into account by the AI system in order to produce a result.

Figure 5 describes some examples of attacks on a generative AI system in an IS.

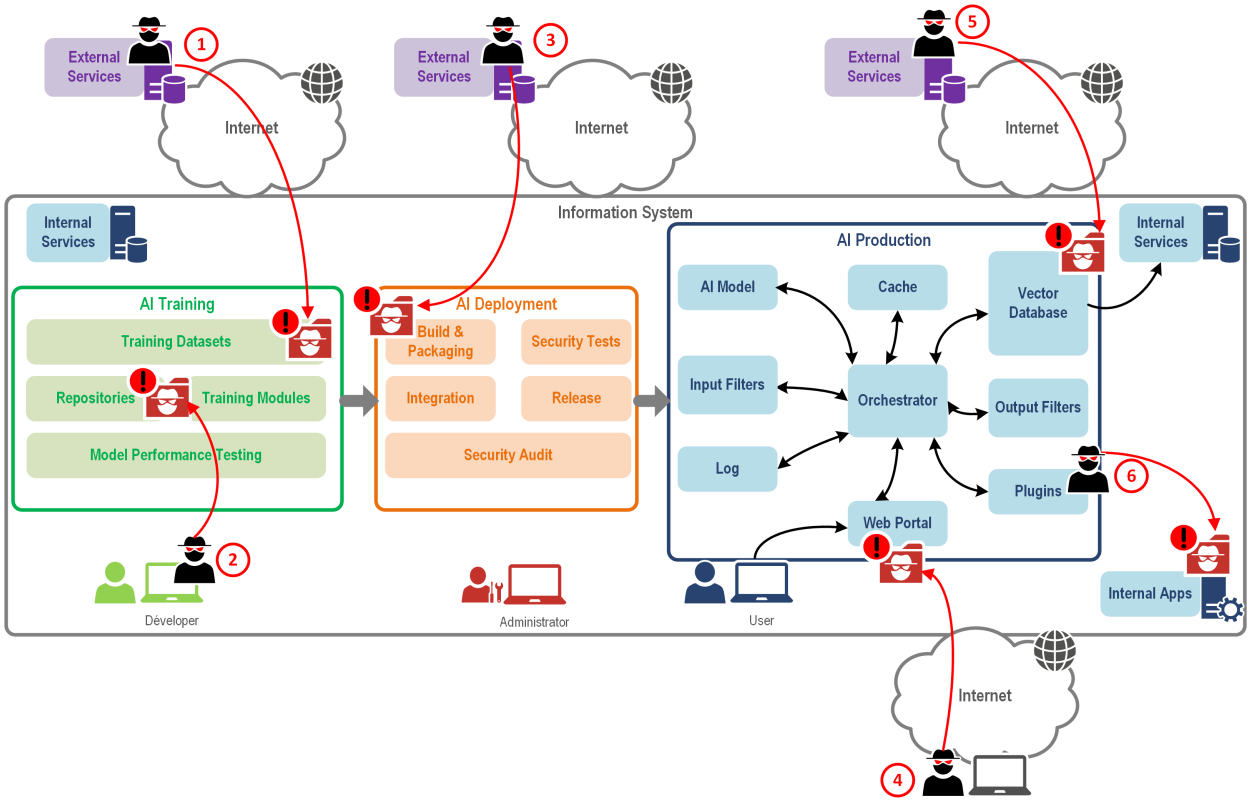

Figure 5 – Attack scenarios on a generative AI system in an IS

1. the attacker has access to a data source and first poisons the data used to train the AI model, which means it can be then used for other purposes once the AI system is in production (e.g. trigger a malicious action based on a specific query);
2. the attacker gains access to the development environment and inserts a backdoor into the code of the AI system (e.g. by directly altering the parameters of the model or the configuration of a technical component of the AI system);
3. the attacker gains access to an external pre-deployment test service and hijacks the integration process (e.g. by sending a misleading or malicious result to the integration chain);
4. either the attacker uses an adversarial attack technique to exfiltrate sensitive data processed by the AI model (e.g. retrieving training data or queries from other users of the service), or uses malicious queries to cause a denial of service;
5. the attacker has access to an external resource accessed by the AI system and sends a malicious response which is integrated by the model (e.g. sending a URL which points to a malicious website controlled by the attacker);
6. the attacker gains access to a plugin used by the AI system and injects malicious commands when performing an action on a business application (e.g. inserting malicious code into the body of an email generated by the AI system).

In the context of this guide (LLM generative AI), the following effects can be identified:

- tarnishing the reputation of services available to the public by altering the proper functioning of generative AI systems (e.g. Chatbot);
- exfiltration of sensitive data from generative AI systems;
- theft of proprietary AI model parameters (weights) [11];
- lateralisation of an attack to other business applications interconnected to generative AI systems (e.g. internal mailing system);
- sabotage of business applications by injecting vulnerabilities into AI-generated source code.

A data leak of an entity's sensitive data is a threat which must be taken into account for all cases, regardless of the use case of the generative AI system. The AI system must incorporate the issue of the access rights and respect information user access restrictions into the responses it provides.

Particular attention should also be paid to indirect attack scenarios involving an AI system, such as the automatic generation of content for another application (insertion of malicious URLs in the response).

Finally, the risk analysis must take into account the responsibility-sharing structure adopted for the project (see figure 2). For example, using a model trained by a third party may give rise to the risk of a supply-chain attack. Untrusted third-party providers can train the model to react in an unexpected manner when it is provided with a particular query. One of the risk reduction measures could be to audit the model itself or not to use it for critical applications.

A risk analysis, carried out using the EBIOS-RM [23] method for example, should therefore start as early as possible, i.e. before the training phase.

R2

Conduct a risk analysis on AI systems before the training phase

Risk analysis of an AI system must address the following issues:

- Map all the elements linked to the AI model: third-party libraries, data sources, interconnected applications, etc.;
- Identify the sub-parts of the AI system that will process the organisation's data, particularly the ones contained in user queries;
- Consider how responsibilities would be shared and how subcontracting would work for each of the phases;
- Identify the direct and indirect effects that incorrect or malicious responses from the AI model to users would have;
- Consider the security of AI model training data.

The recommendations in the next chapter are designed to address these specific threats.

11. Knowledge of model weights can also allow attackers to improve the capability for other attacks.

5 Recommendations

5.1 General recommendations

The use of external libraries and modules must be considered during the design phase of the project, in order to identify potential vulnerabilities associated with these modules. The goal is to provide the maximum level of protection against a supply-chain-attack targeting components required for the proper functioning of the AI system. Please refer to the ANSSI guide on digital risks [22] or the CISA documentation on this subject [8].

R3

Evaluate the level of confidence in the libraries and external modules used in the AI system

It is recommended to map all libraries and external modules used in the project and to assess the level of confidence in these.

In the same way as for software components (libraries and external modules), it is also essential to assess the sources of data not managed by the organisation. These sources could be training data sets retrieved from the Internet, model performance validation sets or additional data sets used during the production phase.

R4

Evaluate the level of confidence in external data sources used in the AI system

It is recommended to map all external data sources used in the project and to assess the level of confidence in these [12].

In general, it is recommended to apply good development practices during the design and implementation of the AI system. These good practices are sometimes grouped together under the name DevSecOps or the term security-by-design. For further information, please refer to the ANSSI guide [28] on this subject, or to the NIST documentation on this subject [10], or follow the recommendations of the NCSC-UK [5] and the CISA [11].

R5

Apply DevSecOps principles to all phases of the project

Secure development best practices should be applied throughout all phases of the project, for example:

- Deploy and secure continuous integration and continuous deployment (CI/CD)

12. The CNIL criteria (https://www.cnil.fr/fr/tenir-compte-de-la-protection-des-donnees-dans-la-collecte-et-la-gestion-des-donnees) or the proposal by *Datasheets for Datasets* (https://arxiv.org/pdf/1803.09010.pdf) can be used to evaluate an external dataset.

chains by applying the least privilege principle for access to the tools in these CI/CD chains;

- Securely manage secrets used in all phases of the project;
- Carry out automated security tests on the source code (static code analysis) and when the source code is executed (dynamic code analysis);
- Protect the integrity of the source code and secure access to it (multi-factor authentication, code signature, access rights, etc.);
- Use secure development languages (fine tuning scripts, model development, maintenance, deployment, etc.).

In order to implement an AI model, the various parameters of this model (weight, bias, etc.) need to be stored in files. Several formats can be used for this purpose, some of which may present a risk of arbitrary code execution, such as those implementing functions for loading serialised objects. It is therefore preferable to use formats that strictly separate the model's data parameters and the model's executable code data.

R6

Use secure AI model formats

We recommend using state-of-the-art security formats, such as *safetensor*. Certain insecure formats, such as *pickle*, should be forbidden.

Generative AI models will have to handle data throughout their lifecycle. Applying confidentiality protection measures to this data can be complicated for the following reasons:

- The data may come from multiple sources and are sometimes combined together in the same set: public data, partner data, internal data, private data, etc. ;
- The volume of data can be very large, particularly in the case of LLM training, which makes processing more difficult;
- Regular updates may be required, particularly when re-training the model;
- Data may need to be pre-processed to lower the level of confidentiality (anonymisation, deletion of fields, etc.);
- Data may be used across different phases of the project: during the model training phase, but also in production when the model needs to access additional data;
- The data can include the AI system's user data during the production phase: data from user queries and the responses provided by the AI model to these users.

It is important to understand that an AI model inherits the sensitivity of the data that contributed to its training and also of the data used to re-train it. An AI model can be vulnerable to a phenomenon known as "regurgitation". In some cases, for example, it may generate responses which are close to the training data, revealing potentially sensitive information.

Depending on the responsibility-sharing scenario adopted (see chapter 2), the data confidentiality issues will be different, and the technical measures to protect against data exfiltration will need to be adapted. For example, if the entity wishes to subcontract the training phase to a service

provider, it is important to ensure that the data stored and processed by this service provider is kept sufficiently confidential (state-of-the-art encryption, isolation of resources from other customers, secure keys used, secure deletion after reallocation of resources, etc.).

Similarly, a proprietary AI model that you want to keep confidential must be subject to specific security measures if it is stored in an untrusted environment (e.g. in a Cloud provider or embedded in physical hosted equipment, for example in the IoT).

R7

Manage data confidentiality issues from the AI system design phase

The project design must map all the data sets used in each phase of the AI system: training (training data sets), deployment (test sets) and production (additional data, vector database, etc.).
This study must include users data from the AI system in production, i.e. user queries and the responses provided by the AI model.
The analysis can also cover confidentiality protection for the parameters of the model itself, for proprietary model, for example.

Access to an AI system also complicates the application of the users rights on the information. There are several categories of data to differentiate here:

- **Training data**: it is not possible to manage user access rights on this data due to the structure of the neural networks;
- **Additional data in production**: it is possible to manage access rights but this depends on the options offered by the tools used (RBAC [13]) to store the information (local data management system, vector database, etc.);
- **User data**: user queries and responses may contain sensitive data. These data are temporarily stored during processing in the AI system and sometimes are used to re-train the model (e.g. alignment with RLHF - Reinforcement learning from human feedback).

The information access rights issue must therefore be considered again each time the model is retrained, including data derived from the use of the model in production (additional business data, user queries, etc.).

R8

Manage users data access rights issue from the AI system design phase

It is important to define the options for the model's structure prior to the project in order to manage the need to know:

- The choice of data used for training (without the ability to manage access rights) and additional data for production (with the ability to manage roles and access rights);
- The model's training strategy, i.e. when is the model re-trained and what underlying data is used for the model's re-training (additional business data, user queries, model responses, etc.).

13. Role Based Access Control

The majority of LLM generative AI have non-deterministic behaviour, and they can also be subject to hallucinations[14]. This uncertainty around the response to a given query means that greater vigilance is required with regard to the indirect consequences of these responses. For example, the interaction of an AI system with other IS resources must not allow critical automated actions to the organisation to be executed.

This precautionary principle must be applied and a generative AI system must not be able to make critical decisions that have a major impact on the business or the protection of assets and people without human intervention (e.g. validation in an HMI). In these particular cases, human capacity for discernment helps to reduce the risk of scenarios that could present a danger to the organisation.

For example, it is important not to use an AI system to automate critical administrative actions on the entity's technical infrastructure (e.g. discovery and automatic deployment of network configurations or firewall rules).

R9

Do not allow AI systems to run automatically critical actions on the IS

An AI system must be configured so that it cannot automatically execute critical IS actions.
These actions may be critical from a business point of view (banking transactions, production of public content, physical impact on humans, etc.) or critical actions on the IS infrastructure (reconfiguration of network components, creation of privileged users, deployment of virtual machines, etc.).

The roles and access rights of AI system developers and administrators must be strictly defined and applied during the project. The principles of secure administration, as described in the ANSSI guide on this subject [21], must be applied in all phases of the AI system's lifecycle.

R10

Manage and secure developer and administrator privileged access to the AI system

All privileged operations on the AI system must comply with secure administration best practices, in particular:

- Privileged operations must be defined and triggering these must be approved: re-training, modification of data sets, new interconnection with an application, change of hosting, etc.;
- Privileged operations must be carried out using dedicated accounts and from a dedicated administration workstation;
- The principle of least privilege must be applied and temporary authentication tokens should be used;
- The development environment must be managed to the same level of security as the production environment.

14. Phenomenon in which a model generates erroneous content that is not based on real data.

The hosting of the AI system, regardless of the phase it is in, must be considered. The level of security must be consistent with the project's security requirements, and especially the confidentiality requirements for the data used in each phase.

In particular, this point must be strictly applied for the model training phase because there are major threats during this phase, as we saw earlier in the chapter 4.

Warning

AI models are considered to have the same level of sensitivity as the data used to design and train them. Rule R9 [12] from the circular "Cloud au Centre" [13] must be applied in the case of French public administration.

R11

Host the AI system in trusted environments consistent with security needs

The hosting of the AI system during the 3 phases of the lifecycle must be consistent with the project's security requirements, especially its confidentiality and integrity requirements. In particular, the security of the model's training data (at rest, in transit, during processing) must not be overlooked.

An AI system's training, deployment and production environments must be siloed separately. This measure reduces the risk of lateralisation between environments. This is particularly important as the groups of people with access to each environment are usually not the same.

R12

Isolate each phase of the AI system into a dedicated environment

The 3 technical environments corresponding to each phase of the AI system's lifecycle should be siloed. This isolation may involve:

- Network isolation: each environment is integrated into a physically or logically dedicated network;
- System isolation: each environment has its own dedicated physical servers or hypervisors;
- Storage isolation: each environment has its own storage hardware or dedicated disks. At the very least, there will be a logical segmentation;
- Accounts and secrets isolation: each environment has its own users and administrators accounts and separate credentials.

For AI systems which are hosted on the Internet, it is recommended to follow the ANSSI recommendations for the design of a secure Internet gateway [25].

R13

Implement a secure Internet gateway for an AI system hosted on the Internet

For AI systems which are hosted on the Internet, it is recommended to follow the isolation best practices of the ANSSI guide on this subject, in particular:

- implement a reverse-proxy function before accessing the AI system web service;

- set up two logical areas for network filtering using firewalls: external filtering on the Internet front end and internal filtering before accessing the AI system;
- do not use any of the entity's internal directories for authentication on the AI system;
- avoid mutualisation of security functions on the same hypervisor in the secure Internet gateway (firewalls, reverse-proxy, logging server, etc.).

If the entity choose a public cloud [15] to host its service and security requirements make this necessary, a SecNumCloud [33] qualified service provider should be chosen.

R14

Prioritise SecNumCloud hosting when deploying an AI system in a public cloud

If the organisation chooses to use public Cloud hosting, it is recommended that a trusted SecNumCloud offer be used in the following cases:

- The data processed by the AI system is considered sensitive;
- The impact of the AI system on the business is considered critical;
- AI system users are not considered trusted.

When designing the project, a downgraded mode must be systematically implemented without any AI in order to meet business needs if the AI system is unavailable or fails.

R15

Provide a downgraded version of business services without an AI system

To prevent malfunctions or inconsistencies in the responses provided by the AI model, it is recommended that, at minimum, there is an AI system bypass procedure for users, in order to meet business needs.

The deployment of generative AI and LLM systems generally involves GPUs [16] to enhance system performance, whether in the training or production phase.

These GPUs may process sensitive data linked to the AI model's operations. To protect against data leaks, these GPU hardware components should be dedicated to the AI system and not be shared with other IS business applications. GPUs, on the other hand, can be shared between several AI models, but only if they have the same level of sensitivity and same security requirements.

R16

Dedicate GPU components to the AI system

Physical GPU components should be dedicated to the processing carried out by the AI system. In the case of virtualisation, hypervisors with access to GPU cards should be dedicated to the AI system, or at least have a hardware filtering function (e.g.: IOMMU [17]) to restrict virtual machine access to the memory on these GPU cards.

15. A public cloud is a hosting service shared between several customers and hosted on the Internet.
16. Graphics processing units

Like most business applications, AI systems can be subject to side-channel attacks. These attacks usually aim to exfiltrate sensitive information or disrupt the proper functioning of AI systems. While most of these attacks are not specific to an AI system, some may nevertheless rely on mechanisms that are specific to generative AI systems.

R17

Manage side-channel attacks on the AI system

It is recommended to ensure that the AI system is not vulnerable to attacks via auxiliary channels (time, energy consumption, etc.) which could, for example, enable an attacker to rebuild a response provided by an AI model.

5.2 Recommendations for the training phase

The issue of data confidentiality has already been covered in a previous general recommendation (see R7). In particular, and regarding the number of vulnerabilities published on generative AI tools, it should be assumed that a user with access to a trained AI model could potentially have access to the training data for this model.

To reduce the risks associated with the confidentiality of training data, it is sometimes necessary to use an anonymisation process or generate a synthetic dataset from the original raw data. In some cases, these measures can resolve the issue of protecting information, but it is nevertheless important to be vigilant about attacks aimed at retrieving the initial information from anonymised or synthetic data [18]: attacks by attribute or membership inference, re-identification based on cross-referencing with other datasets, etc.

R18

Train an AI model only with data which users can legitimately access

It is strongly recommended to train a model with data of a level of sensitivity consistent with the users' access rights.

As we saw earlier, attacks specifically targeting the training phase of a model are possible, such as the injection of malicious data into training datasets, or modifying certain data to cause the model to malfunction once it has been deployed in production.

R19

Protect the integrity of AI model training data

The integrity of the model's training data should be ensured throughout the training cycle. This protection may take the form of systematic checking the signature or hash of the files used (or compressed archives with all this data).

17. Input–output memory management unit
18. see, for example, https://cdn.arstechnica.net/wp-content/uploads/2024/03/LLM-Side-Channel.pdf
18. Refer to the CNIL's work on this subject: https://linc.cnil.fr/donnees-synthetiques-et-lhomme-crea-les-donnees-son-image-22.

R20

Protect the integrity of AI system files

The integrity of the trained model files should be protected and regular checks done to ensure that these have not been altered. This recommendation also applies by extension to all the files inherent to the proper working of the AI system (scripts, binaries, etc.).

In the majority of use cases, a trained AI model does not need to be subject to regular modifications or adjustment of its parameters. If a malfunction is detected, or when optimising the model's performance, it is preferable to carry out re-training operations using the dedicated training environment.

To respect this, continuous learning methods, also known as online learning (where the model learns in real-time from the data sent as input), should be avoided as far as possible. Using offline learning methods, based on selected and tested data sets, reduces the risk of the model malfunctioning or being poisoned.

An AI model can be re-trained on a recurring and fixed basis (e.g. every month), triggered when a performance gap crosses a given threshold or when the data used to train the model are no longer relevant, or on demand on an ad-hoc basis.

R21

Do not re-train an AI model in production

It is strongly recommended not to re-train an AI model directly in production. Re-training should start with the 3-phases cycle, in the suitable environments for each phase.

5.3 Recommendations for the deployment phase

The deployment of a generative AI system must be based on a secure deployment environment based, for example, on fully harnessed and robust CI/CD chains.

These CI/CD chains must be operated from an administration system and from dedicated, robust administrator workstations.

R22

Secure the production deployment chain for AI systems

It is recommended that generative AI systems be deployed from an administrative IS, in compliance with ANSSI's secure administration best practices guide [21].

A security audit should be carried out by specialised teams which are trained in the specificities of AI systems. This phase must take place before deployment to production in order to test the vulnerabilities inherent to AI systems (adversarial attacks, etc.).

R23

Conduct security audits of AI systems before deployment to production

Robustness and security tests of AI systems are recommended. These tests can be:

- Standard penetration tests on the usual technical components of an AI system: web servers, orchestrator, database, etc. ;
- Security tests on developments made in the AI system (using SAST or DAST tools, for example);
- Automated tests [19]specifically targeting vulnerabilities related to AI models (adversarial attacks, model extraction, etc.);
- Manual auditor tests specifically aimed at testing the robustness of a generative AI model in more sophisticated attack scenarios.

To carry out safety audits on a generative AI system, it is possible to use ANSSI-qualified PASSI [31] service providers.

R24

Conduct business tests of AI systems before deployment to production

Performance and quality tests should be carried out on the answers provided by a generative AI system.

Information

Functional testing of the AI system can take place continuously at a given frequency and not only during deployment. This can be used to detect model malfunctions at an early stage and correct them more reactively.

5.4 Recommendations for the production phase

As mentioned earlier, it is difficult to apply the restrictions for user right access to the training data of a model, which may be subject to attacks aiming to extract this data by querying the model (exfiltration).

Similarly, some malicious requests may be aimed at hijacking the generative AI service, for example by inducing hallucinations or incorrect responses.

As part of a defence-in-depth approach, it is important to explore the possibility of detecting or blocking some malicious queries intented, for example, to extract the model data or additional data (this additional data may include user inputs in certain cases).

This protection can also be useful in reducing the risk of the model leak. If the model has been trained on sensitive data, the model parameters leak can lead to some of this data leak as a result of some attacks (e.g. model inversion attack or membership inference attack). As such, responses to users must be as simple as possible (character strings only) and must not return any prediction score or other internal model mechanisms.

19. There are a number of specialised tools available, such as https://github.com/microsoft/responsible-ai-toolbox, https://github.com/Trusted-AI/adversarial-robustness-toolbox, or https://github.com/protectai/ai-exploits.

Finally, depending on the use case, it may be appropriate to define a limit to the size of the answers provided by the AI model. This can reduce the risk of data leak.

R25

Protect the AI system by filtering user input and output

These functions should be implemented to protect against data leak or model leak in responses:

- A function to filter malicious user queries before these are sent to the model;
- A filter function for queries deemed to be non-legitimate from a business point of view;
- A filter function for internal model information (parameters, training) in the responses;
- A filter function for information defined as sensitive in responses (e.g. private details, project references, etc.);
- A limit on the size of responses (maximum number of characters).

The AI system's interaction with other business applications or other IS technical resources can be source of vulnerabilities.

These interactions often take the form of plugins offered by AI model editors. These plugins will enable the AI system to be interconnected with bureautic tools and social networks, or potentially critical infrastructure components (identity management, network resources, etc.).

These interactions can also facilitate the lateralisation of an attacker on the IS, if he takes advantage of a vulnerability in the AI system.

The literature on this topic often refers to the risks of *indirect prompt injection* and the problems that may result from sending uncontrolled data to a generative AI model [20] (for example, the content of a received email or a web page resulting from a search). This kind of use is more problematic when an action is carried out without human validation (see recommendation R9).

It is therefore essential to be able to control the interaction of the AI system with other IS resources.

R26

Manage and secure the interactions of the AI system with other business applications

All the interactions and network flows of the AI system must be documented and approved. Network flows between the AI system and other resources must comply with the state-of-the-art in terms of security:

- They must be strictly filtered at network level, encrypted and authenticated (e.g. by following the ANSSI TLS [19] guide);
- They must use secure protocols (e.g. OpenID Connect) when using an identity provider [24];

20. see the article "Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection" for more details: https://arxiv.org/abs/2302.12173

- In addition to authentication, the authorisation of access to the resource must also be checked;
- They must be logged at the appropriate level of granularity.

R27

Limit automatic actions performed by an AI system handling uncontrolled inputs

It is strongly recommended that automated actions on the IS be limited or even prohibited when these are triggered by an AI system and uncontrolled inputs (e.g. data from the Internet or emails, etc.).

Depending on the use case of the AI system and its criticity from a business point of view, it may be appropriate to deploy it in one or more dedicated environments, not shared with other business applications.

R28

Isolate the AI system in one or more dedicated technical environments

It is recommended that the AI system be siloed into dedicated logical zones, in order to limit the risk of an attacker who has compromised the system moving laterally.

Actions on the AI system must be logged with adequate level of information granularity, in particular regarding the inputs and outputs of the AI model.

For the purposes of traceability and explicability of the AI system, it is important to make a clear distinction between the queries made by users and the data actually sent to the AI model. In fact, for performance and security reasons, user queries may be subject to specific pre-processing and formatting before being sent to the model.

These two pieces of information are crucial in facilitating the management of an incident and must be traceable in the AI system's application logs. The aim is to be able to fully rebuild an event on the AI system, when detecting a malicious query, for example.

Regarding the architecture of a logging system, refer to the ANSSI's general guide on this subject [27].

R29

Record all processing carried out within the AI system

All processing carried out on the AI system should be logged at the correct level of granularity, in particular:

- User queries (ensuring these are secured if they contain sensitive data);
- The input processing carried out on this query before it is sent to the model;
- Calls to plugins;
- Calls for additional data;
- The processing carried out by the output filters;

- Responses to users.

Warning

User data logging must comply with CNIL [9] requirements concerning the protection of personal data (GDPR) and in particular for how long is this data kept on the AI system.

5.5 Special case of AI-assisted source code generation

Generative AI tools can be specialised and specifically trained to generate source code in a lot of programming languages.

These resources can help developers save time, but there are also risks regarding the quality of the code (pushing vulnerabilities) or the insertion of backdoors if an attacker has compromised the AI model.

It is therefore important to be careful about the source code generated by AI.

R30

Check AI-generated source code systematically

The source code generated by AI must be subject to security measures to ensure that it is not harmful:

- Prohibition of automatic execution of AI-generated source code in the development environment;
- Prohibition of automatic commit of AI-generated source code to repositories;
- Integration of an AI-generated source code remediation tool [3, 4] into the development environment;
- Check that the libraries referenced in the result of the source code generated by AI are harmless;
- The quality of the source code generated from sufficiently sophisticated standard queries needs to be checked regularly by a human.

R31

Limit AI source code generation for critical application modules

It is strongly recommended that generative AI tools are not used to generate blocks [21] of source code intended for critical modules :

- Cryptography modules (authentication, encryption, signatures, etc.);
- User and administrator access rights management modules;
- Sensitive data processing modules.

.

21. A block here refers to a complete set of instructions in the source code, for example the complete definition of a function, a procedure, an object class, a shell script, etc.

R32

Raise developers awareness of the risks associated with AI-generated source code

Awareness-raising campaigns on the risks associated with the use of AI-generated source code should be carried out. This awareness can be based on public reports on the subject or research papers [22] demonstrating the presence of vulnerabilities in AI-generated code.
In addition, developers can also be trained in AI tools to optimise their queries (prompt engineering [23]) to improve the quality and security of the generated code.

Information

Depending on the use case, it may also be appropriate to specifically train a model (alignment stage) so that it cannot generate deliberately malicious code.

5.6 Special case of consumer AI services hosted on the Internet

If the entity wishes to offer a service based on generative AI to the public, specific care must be taken to ensure the security of this service, given its high exposure.

An additional threat to consider during the risk analysis is the potential of damage to the entity's reputation.

R33

Strengthen security measures for AI services hosted on the Internet

Specific attention should be paid to certain security measures for services available to the public, in particular:

- Training the AI model using only publicly available data;
- Ensuring that users of the AI system have been authenticated before use;
- Systematic analysis of user queries on the AI system;
- Checking and validating responses before these are sent to users;
- Protecting the confidentiality of user data (history of queries and responses, etc.);
- Implementing measures against distributed denial of service (DDoS) [18] attacks;
- Securing the web service at the user front end [26].

5.7 Special case of using third-party generative AI solutions

In this final section, the guide goes over the case where the entity is not managing a generative AI service but is a customer of a third-party generative AI service (see responsibility-sharing scenarios

23. The reports by *Snyk* (https://snyk.io/fr/reports/ai-code-security/) could be cited as an example, or the research carried out by Stanford University on this subject (https://arxiv.org/pdf/2211.03622.pdf).
24. For example, an initial AI code generation query can be combined with by a static analysis test of this code and then a second query asking the AI to treat the vulnerabilities detected in the original generated code.

in chapter 2). The purpose of this last section is to restate the essential points which must be observed and taken into account by users of these third-party services.

Because they are so easy to use, it's tempting to use generative AI tools that are available on the Internet to process business data, for example for text translation. Sending information (text, images, documents) to a generative AI service which is available to the public is the same as pushing the information on a storage space belonging to editors.

Isolation between clients and protection of the confidentiality of data sent to an Internet AI system are not always in the state-of-the-art and rely solely on trust in the service provider. In this regard, it is important to note that with the majority of services, the data sent to the service is collected and used by the service provider to optimise the models [24].

Therefore, sensitive data must absolutely not be sent to third-party generative AI services such as *ChatGPT*, *Gemini*, *Copilot*, *DeepL* (text translation) or *Perplexity*, to name only the most popular services. The relevant data includes:

- French *Diffusion Restreinte* data [30] or classified data [32];
- Research work relating to French PPST [20];
- Personal data (private information, contact details, etc.);
- The company's contractual, legal and financial data;
- IT secrets, such as passwords or authentication tokens (API keys).

R34

Do not use generative AI tools on the Internet for professional use involving sensitive data

As the client entity does not control the generative AI service, it is impossible to ensure that the confidentiality of data submitted for input meets the entity's security requirements.
As a precautionary measure, it is therefore essential never to include any of the entity's sensitive data in user queries.

Warning

This recommendation also concerns the use of generative AI tools to generate synthetic datasets for training or fine-tuning an AI model.

Some third-party generative AI tools offer connections with office automation tools or standard business applications on the Internet. Particular attention must be paid to the configuration of generative AI tools' rights of access to the entity's business data: emails, documents areas, source code repositories, audio and video conferencing services, etc.

24. see *ChatGPT*'s usage policy for example: https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed

R35

Perform regular reviews of the configuration of rights for generative AI tools on business applications

A review of access rights for generative AI tools should be carried out as soon as the entity activates the product to ensure that the rights set by default are not too low or open by design.
Finally, access rights must be reviewed on a regular basis (e.g. every month), to ensure that the product's functional and security updates have no impact on user information access rights.

Recommendation List

Bibliography

[1] *BSI - Artificial Intelligence*.
Site institutionnel, BSI.
https://www.bsi.bund.de/EN/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/kuenstliche-intelligenz_node.html.

[2] *CNIL - Intelligence artificielle (IA)*.
Site institutionnel, CNIL.
https://www.cnil.fr/fr/intelligence-artificielle-ia.

[3] *NIST - Source Code Security Analyzers*.
Site institutionnel, NIST.
https://www.nist.gov/itl/ssd/software-quality-group/source-code-security-analyzers.

[4] *OWASP - Source Code Analysis Tools*.
Technical report, OWASP.
https://owasp.org/www-community/Source_Code_Analysis_Tools.

[5] *NCSC-UK - Secure development and deployment guidance*.
Site institutionnel, NCSC-UK, novembre 2018.
https://www.ncsc.gov.uk/collection/developers-collection.

[6] *ENISA - Artificial Intelligence Cybersecurity Challenges*.
Site institutionnel, ENISA, décembre 2020.
https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges.

[7] *ENISA - Securing Machine Learning Algorithms*.
Site institutionnel, ENISA, décembre 2021.
https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms.

[8] *CISA - Securing the software supply chain*.
Site institutionnel, CISA, août 2022.
https://media.defense.gov/2022/Sep/01/2003068942/-1/-1/0/ESF_SECURING_THE_SOFTWARE_SUPPLY_CHAIN_DEVELOPERS.PDF.

[9] *CNIL - IA : comment être en conformité avec le RGPD ?*
Site institutionnel, CNIL, avril 2022.
https://www.cnil.fr/fr/intelligence-artificielle/ia-comment-etre-en-conformite-avec-le-rgpd.

[10] *NIST - Secure Software Development Framework (SSDF) Version 1.1: Recommendations for Mitigating the Risk of Software Vulnerabilities*.
Site institutionnel, NIST, février 2022.
https://csrc.nist.gov/pubs/sp/800/218/final.

[11] *CISA - Defending Continuous Integration/Continuous Delivery (CI/CD) Environments*.
Site institutionnel, CISA, juin 2023.
https://media.defense.gov/2023/Jun/28/2003249466/-1/-1/0/CSI_DEFENDING_CI_CD_ENVIRONMENTS.PDF.

[12] *DINUM - Le Cloud pour les administrations*.
Site institutionnel, DINUM, Mai 2023.
https://www.numerique.gouv.fr/services/cloud/regles-doctrine/#contenu.

[13] *Doctrine d'utilisation de l'informatique en nuage par l'État - Cloud au centre*.
Site institutionnel, LEGIFRANCE, Mai 2023.
https://www.legifrance.gouv.fr/download/pdf/circ?id=45446.

[14] *NCSC-UK - Guidelines for secure AI system development*.
Site institutionnel, NCSC-UK, novembre 2023.
https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development.

[15] *NIST - Artificial Intelligence Risk Management Framework*.
Site institutionnel, NIST, janvier 2023.
https://www.nist.gov/itl/ai-risk-management-framework.

[16] *NIST - Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*.
Site institutionnel, NIST, janvier 2024.
https://csrc.nist.gov/pubs/ai/100/2/e2023/final.

[17] *Guideline for a healthy information system*.
Guide ANSSI-GP-042-EN v2.0, ANSSI, septembre 2017.
https://cyber.gouv.fr/en/publications/guideline-healthy-information-system-42-measures.

[18] *Comprendre et anticiper les attaques DDoS*.
Guide Version 1.0, ANSSI, mars 2015.
https://cyber.gouv.fr/guide-ddos.

[19] *Security recommendations for TLS*.
Guide SDE-NT-035-EN v1.1, ANSSI, janvier 2017.
https://cyber.gouv.fr/en/publications/security-recommendations-tls.

[20] *Protection du potentiel scientifique et technique de la nation*.
Guide ANSSI-PA-049 v1.0, ANSSI, avril 2018.
https://cyber.gouv.fr/guide-zrr.

[21] *Recommendations to secure administration of IT systems*.
Guide ANSSI-PA-022-EN v2.0, ANSSI, avril 2018.
https://cyber.gouv.fr/en/publications/recommendations-secure-administration-it-systems.

[22] *Controlling the digital risk - The trust advantage*.
Guide ANSSI-PA-070-EN v1.0, ANSSI, novembre 2019.
https://cyber.gouv.fr/en/publications/controlling-digital-risk-trust-advantage.

[23] *EBIOS Risk Manager - Methodological sheets*.
Guide ANSSI-PA-058-EN v1.0, ANSSI, novembre 2019.
https://cyber.gouv.fr/en/digital-risk-management.

[24] *Recommandations pour la sécurisation de la mise en œuvre du protocole OpenID Connect*.
Guide ANSSI-PA-080 v1.0, ANSSI, septembre 2020.
https://cyber.gouv.fr/guide-oidc.

[25] *Recommandations relatives à l'interconnexion d'un système d'information à Internet*.
Guide ANSSI-PA-066 v3.0, ANSSI, juin 2020.
https://cyber.gouv.fr/guide-interconnexion-si-internet.

[26] *Recommandations pour la mise en œuvre d'un site Web : maîtriser les standards de sécurité côté navigateur*.
Guide ANSSI-PA-009 v2.1, ANSSI, avril 2021.
https://cyber.gouv.fr/guide-sites-web.

[27] *Recommandations de sécurité pour l'architecture d'un système de journalisation*.
Guide DAT-PA-012 v2.0, ANSSI, janvier 2022.
https://cyber.gouv.fr/guide-journalisation.

[28] *Les essentiels - DevSecOps*.
Guide Version 1.0, ANSSI, février 2024.
https://cyber.gouv.fr/publications/devsecops.

[29] *Recommandations de sécurité pour un système d'IA générative*.
Guide ANSSI-PA-102 v1.0, ANSSI, avril 2024.
https://cyber.gouv.fr/guide-ia-generative.

[30] *Instruction interministérielle n°901*.
Référentiel Version 1.0, ANSSI, janvier 2015.
https://cyber.gouv.fr/ii901.

[31] *Prestataires d'audit de la sécurité des systèmes d'information. Référentiel d'exigences*.
Référentiel Version 2.1, ANSSI, octobre 2015.
https://cyber.gouv.fr/referentiels-dexigences-pour-la-qualification.

[32] *Instruction générale interministérielle n°1300*.
Référentiel, SGDSN, août 2021.
https://cyber.gouv.fr/igi1300.

[33] *Prestataires de services d'informatique en nuage (SecNumCloud). Référentiel d'exigences*.
Référentiel Version 3.2, ANSSI, mars 2022.
https://cyber.gouv.fr/secnumcloud.

Version 1.0 - 29/04/2024 - ANSSI-PA-102
Licence ouverte / Open Licence (Étalab - v2.0)
ISBN : 978-2-11-167169-0 (papier)
ISBN : 978-2-11-167170-6 (numérique)
Dépôt légal : September 2024

AGENCE NATIONALE DE LA SÉCURITÉ DES SYSTÈMES D'INFORMATION

ANSSI - 51 boulevard de La Tour-Maubourg, 75700 PARIS 07 SP
cyber.gouv.fr / conseil.technique@ssi.gouv.fr

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité