What is the scope of the AI how-to sheets?

07 juin 2024

The CNIL wishes to provide concrete clarifications and recommendations for the development of artificial intelligence (AI) systems and the creation of datasets used for their training, when they involve personal data. The CNIL defines the scope of the different how-to sheets.

Item detail:

The following sheets **concern only the development phase of AI systems, not their deployment,** where this involves the processing of personal data. They are limited to the processing of data subject to the General Data Protection Regulation (GDPR).

They aim at accompanying a large number of professionals with both legal and technical profiles (such as data protection officers, legal officers, legal professionals, AI practitioners, etc.).

Please note:
These how-to sheets, adopted following a public consultation, provide a framework to help professionals with **their compliance**. **They recall the obligations imposed by GDPR and make recommendations to comply with them**. **These recommendations are not binding**: **data controllers may deviate from it**, under their responsibility and **provided that they can justify their choices**. **Some recommendations are also made as good practices** and allow one to go beyond GDPR obligations.

The presence of personal data

The how-to sheets relate to dataset creation activities and their use in the development of AI systems where all or part of that data is personal data. In practice, three cases may be encountered:

- **It is certain that no personal data is present in the dataset**: the how-to sheets are not applicable (although some recommendations may be relevant as good practices).

- **It is certain that personal data are present:** the how-to sheets apply. This is the case for AI systems developed from videos or images of people, voice recordings, structured personal data, etc. It should be noted that the European texts lay down the rule that 'mixed' datasets are governed by the GDPR, if both types of data are inextricably linked.

- **Personal data may be present**: this is a frequent case for which the collection of personal data is not expressly desired. For example:
    - residual presence of persons or license plates in images;
    - occurrences of surnames, first names, addresses, etc. in textual data such as comments or prompts, etc.

      In the latter case, the how-to sheets apply. However, verification operations may be carried out as a result of the collection in order to delete the remaining personal data. This can be achieved:

    - **by manual verification**, e.g. when annotating the data;
    - **by automatic verification**, e.g. by using techniques for detecting persons/faces in images, by nameentity recognition methods (NER), etc.

In such cases, provided that the original personal data has been anonymized, and for processing operations subsequent to such deletion, the data loses its personal character and the application of the how-to sheets is no longer mandatory. Issues related to the risks associated with the use of the system will be the subject of how-to sheets published at a later date.

The AI systems concerned

The CNIL's how-to sheets concern the development of systems implementing artificial intelligence techniques involving the processing of personal data. These are referred to as 'AI systems'.

The definition of AI systems covered by these how-to sheets is aligned with the definition of the EU AI Act just adopted (see box below): a*n AI system is a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments*.

In practice, the AI systems concerned include systems based on machine learning (supervised, unsupervised, by reinforcement) and systems based on logic and knowledge (knowledge-based, inference and deduction engines, symbolic reasoning, expert systems, etc.), as well as hybrid approaches.

The how-to sheets concern those systems whether operational use in the deployment phase is defined at the development stage, or whether they are general purpose AI systems (GPAI), for example relying on "foundation" models, because of their ability to be reused and adapted for different applications and use cases.

They also cover all AI systems as defined above, whether training is done "once and for all" or continuously. In the case of continuous learning, the data collected when the system is deployed is reused for the system's iterative improvement.

Finally, the how-to sheets also concern the process of training existing AI models (fine-tuning / transfer learning), regardless of their integration into a system itself, where they involve personal data.

The phases of the treatment concerned

As illustrated in the previous diagram, the establishment of an AI system based on machine learning requires, in principle, the succession of two distinct phases:

- **the development phase**: it consists of designing, developing and training an AI system.
- **the deployment phase**: it consists of putting into use the AI system developed in the first phase.

The development phase includes all stages from the development of the AI system to its deployment (production phase), namely:

- **the design of the system**: choice of the architecture, including the learning method(s) and, where appropriate, the selection of pre-trained models, identification of the necessary data and first tests, pilots;
- **the establishment of the database**: collection and pre-processing (cleaning, annotation, feature extraction, data allocation);
- **learning**: model training, possible fine tuning, adjustment and validation of hyperparameters, performance tests;
- **and sometimes integration**: when the final product expected at the end of the development phase is a system and not a model (for instance a general purpose AI system), inserting the trained model into the information system, connecting to other software components, developing a user interface, writing a user documentation, etc.

In many cases, the development of an AI system will be based on fine-tuning or transfer learning. The CNIL considers that this phase constitutes a second phase of development, distinct from that which produced the original model.

In the case of continuous learning, data is collected during the deployment and development phases (through the use in production of the system), for future improvement. Continuous training is therefore included in this development phase via a feedback loop.

It should be noted that in addition to these two phases, there is a phase of shutdown of the AI system or deletion of personal data that it contained. These sheets do not specify the operations to be carried out during this phase, which are also covered by the regulations on the protection of personal data.

Applicable texts

Regulations on the protection of personal data

The how-to sheets concern, in particular, use cases relating to the phase of development of an AI system (scientific research, research and development, personalisation of a commercial product, improvement of the public service provided to the user, etc.) for which the GDPR is applicable.

Reminder on the scope of the GDPR

The GDPR applies to any organisation:

- public and private, regardless of size (company, administration, community, association, etc.);
- who processes personal data, on its behalf or not;
- established in the territory of the European Union where the processing is carried out in the course of the activities of one of its establishments in the territory of the Union, irrespective of whether or not the processing takes place in the Union;
- or who, not established in the territory of the European Union, directly targets natural persons in the European Union or monitors their behaviour.

Examples:

- **Application of the GDPR to the re-use of datasets created outside the EU**: the GDPR is applicable to the re-use of datasets by a controller or processor established in the European Union where the processing is carried out in the context of the activities of one of its establishments in the territory of the Union, even if those databases have been set up outside the EU and contain the personal data of persons outside the EU. In this case, the controller is therefore obliged to comply with the applicable data protection regulations.

- **Application of the GDPR to the reuse of models trained outside the EU:** the GDPR applies to the use of models created outside the European Union by a controller or processor established within the European Union where they contain personal data and the processing is carried out in the context of the activities of one of its establishments in the territory of the Union.

Data processing in the development phase of the AI system subject to the scope of the Law enforcement directive as well as those relevant to State security and national defense are therefore excluded from the scope of these how-to sheets. However, the recommendations provided can serve as inspiration.

Other applicable regulations

If these how-to sheets are intended to clarify how the development of AI systems can comply with personal data protection obligations, other regulations, which are not addressed directly, may apply. This is, for example, the case of rules on intellectual property law or the regulation on data governance (DGA), which regulates data intermediation services or data altruism.

Others are not yet applicable. This is particularly the case with the proposal for a European AI regulation which aims to frame the development and deployment of AI systems within the European Union.

Finally, sector-specific regulations apply to AI systems developed or deployed for certain applications subject to specific regulation (health, finance, critical systems, etc.). It is up to each controller to determine the applicable regulations and to turn to the relevant regulators.

Interplay between the how-to sheets and the EU AI Act

The European AI Act distinguishes several categories of systems according to their level of risk with regard to product safety and fundamental rights: prohibited systems, high-risk systems, systems requiring transparency guarantees and minimum risk systems. It thus provides for different degrees of obligations based mainly on AI system providers.

CNIL's how-to sheets have been drawn up with a view to an intelligible interplay with these future obligations (e.g. in terms of qualification of actors and risk assessment).

It should be noted, however, that these how-to sheets apply, under constant law, to any processing of data subject to the GDPR in the context of the development of an AI model or system, regardless of the entry into application of the European rules on artificial intelligence. The CNIL also recalls that the AI Act is not intended to replace data protection obligations but to complement them.

The elaboration of more precise rules on the articulation between these different requirements is the subject of European work (within the European data protection board) in which the CNIL actively participates and which will lead to subsequent publications.

Table of content

Next : Determining the applicable legal regime >

Determining the applicable legal regime

07 juin 2024

When it contains personal data, the creation of training datasets, as well as the development phase, must comply with the relevant regulations. The CNIL helps you determine the legal regime applicable to data processing in the development phase.

The principle

The development and deployment phases of an AI system constitute separate processing of personal data, subject to personal data protection regulations. There are different legal regimes depending on the processing:

- **the regime resulting from the General Data Protection Regulation (GDPR)** which is intended to apply to all processing of personal data, both in the public and private sectors, with the exception of processing that fall under the following two specific schemes;
- **the specific regime for the law enforcement sectors** (Title III of the French Data Protection Act);
- **the regime concerning the national defence or the security of the State** governed by the provisions of the French Data Protection Act.

The purpose of this how-to sheet is to define the cases where the data processing in the development phase is subject to the same legal regime as the data processing in the deployment phase, and the cases where they are subject to separate regimes.

As a reminder: the principles and recommendations formulated in the following how-to sheets concern only processing that fall within the scope of the GDPR.

Find out more:

- The scope of CNIL's AI how-to sheets;
- The scope of the Law enforcement Directive.

In practice

In order to determine the regime for data processing in the development phase, two cases must be distinguished.

Case 1: the operational use of the AI system in the deployment phase is defined from the development phase

In the event that the operational use of the AI system in the deployment phase is identified as early as development phase and if the processing operations implemented during the development phase pursue exclusively the same purpose as those in the deployment phase, it may be considered that they are generally covered by the same legal regime (see Conseil d'Etat, 22 July 2022, N° 451653).

This will in particular be the case where the choice of the development of a specific AI system is one of the means identified to achieve the purpose set for the system to be deployed.

It should be noted that the 'law enforcement' regime (Title III of the French Data Protection Act) may apply to processing in the development phase if the following conditions are met:

- the operational use of the AI system is identified from the development stage, so that the processing operations implemented during the development phase have the same purpose only as those in the deployment phase;

- the use of the AI system exclusively pursues the purposes of preventing, detecting, investigating and prosecuting criminal offences or executing criminal penalties, including the protection against and the prevention of threats to public security;

- the controller in development is a 'competent authority'.

Case 2: the operational use of the AI system in the deployment phase is not clearly defined in the development phase (general purpose AI system)

The development phase and deployment phase of the AI system can be decorrelated.

It is not always possible to clearly identify the purpose of the processing in the deployment phase from the development phase. Some AI systems (*general purpose AI systems*) are developed without a specific operational use and then eventually operated in a second stage.

The legal regime of the development phase is therefore not systematically the same as the one determined in the deployment phase.

In this case, it is generally considered, depending on a case-by-case analysis, that processing in the development phase is subject to the GDPR.

Example: an organisation wishes to develop a voice recognition model capable of identifying a speaker and his/her language in order to commercialise it for different operational uses in the production phase (e.g. tools for identifying people by voice assistants or voice translation applications on a mobile device, etc.).

In this case, the creation of the training dataset falls within the scope of the GDPR.

This does not rule out, depending on the operational use of the AI system, that the processing during the deployment phase may be subject to the ‘law enforcement’ regime, if it is carried out by a competent authority for the purposes of the prevention, detection, investigation and prosecution of criminal offences or the execution of criminal penalties.

Example: a company develops an image classification system to detect the crossing of an area. It then commercialises it to several entities:

- in the first case, it sells it to a company that uses it for statistical purposes to measure the influx of people entering a mall. In this case, processing in the development and deployment phase will be subject to the GDPR.

- in the second case, it sells it to a national police service that uses it to detect persons crossing prohibited areas for prosecution. In this case, the processing in the development phase will be subject to the GDPR, but processing in the deployment phase will be subject to the ’law enforcement’ regime.

< Previous : What is the scope of the AI how-to sheets ?

Table of content

Next : Defining a purpose >

Defining a purpose

07 juin 2024

The creation of a training dataset containing personal data is a processing of personal data which, pursuant to the GDPR, must have a purpose that is ’specified, explicit and legitimate’. The CNIL helps you define the purpose taking into account the specificities of AI systems development

The principle

The purpose of the processing is the aim of the use of personal data. This objective must be specified, i.e. defined as soon as the project starts. It must also be explicit, that is to say, known and understandable. Finally, it must be legitimate, i.e. compatible with the tasks of the organisation.

The data must not be further processed in a manner incompatible with this initial purpose: the principle of purpose limitation restricts how the controller may use or reuse these data in the future.

The requirement of a specified, explicit and legitimate purpose is particularly important, as it determines the application of other principles of the GDPR, including:

- **the principle of transparency:** the purpose of the processing must be brought to the attention of the data subjects so that they are able to know the reason for the collection of the data concerning them and to understand the use that will be made of it;
- **the principle of data minimisation:** the data selected must be adequate, relevant and limited to what is necessary for the purposes for which they are processed;
- **the principle of storage limitation:** the data may only be kept for a limited period, defined according to the purpose for which it was collected.

Find out more: Defining a purpose

How to define the purpose of the processing when the operational use is identified from the development stage?

How to define the purpose of processing for the development of general purpose AI systems?

How to define the purposes of the development of an AI system for scientific research ?

< Previous : Determining the applicable legal regime

Table of content

Next : Determining the legal qualification of AI system providers >

Determining the legal qualification of AI system providers

07 juin 2024

AI system providers who intend to create training datasets with personal data must determine their qualification under the GDPR: they may be qualified as controllers, joint controllers or processors.

Several operators may intervene in the development of an AI system, with varying degrees of involvement in the processing of personal data. In particular, there are:

- AI system providers developing or delegating the development of a system and placing it on the market or putting it into service under their own name or brand, against payment or for free.
- importers, distributors, and users of these systems (i.e. those deploying AI systems).

The qualification of the operators involved in each processing, within the meaning of the GDPR, must be analysed on a case-by-case basis.

The controller

The principle

The controller is the natural or legal person who determines the purposes and means of the processing, i.e. who decides on the “why” and “how” of the use of personal data.

The essential means of processing are those which are closely related to the purpose and scope of the processing, such as the type of personal data collected, the hardware and software used for the processing as well as their security, the duration of the processing, the categories of recipients and the categories of data subjects.

In practice

Some clues may help to conduct the analysis on a case-by-case basis to determine who is responsible for the processing.

A provider who is **at the initiative** of the development of an AI system and who **creates the training dataset based** on data that it has selected on its own account, may be qualified as **a controller.**

The same applies to a supplier who entrusts the creation of such a base to a service provider through sufficiently detailed documented instructions (see the role of processor below).

It should be noted that in some cases a provider will have recourse to a third-party who has already created a dataset as the controller (on its own initiative). It will then be necessary to identify the processing for which the provider is responsible, such as the re-use on its own account of a dataset already constituted.

Examples of controllers:

1. A video streaming platform wants to develop a recommendation AI system. For this purpose, it reuses a dataset of its customers that was originally collected for the purpose of providing the service.
   The streaming platform that creates the training dataset is responsible for this new processing since it has decided on the purpose (train a recommendation AI system) and the essential means of processing (i.e. the dataset it has already collected for another purpose).
2. The provider of a conversational agent who trains its large language model (LLM) from publicly available data on the Internet is controller of the reuse of publicly available personal data on the Internet. Indeed, the provider decides both the purpose (proposing a conversational agent) and the essential means of processing (selecting the data to be re-used).
3. A provider develops an AI system based on a pre-trained model with personal data. The provider intends to retrain or adjust the model (through fine-tuning or transfer learning) with a dataset that it set up, at its initiative. In such a case, that provider will have to be classified as a controller, provided that it pursues a purpose of its own and for which it determines itself the essential means.
4.

Reuse of data collected by another organisation

When the provider trains its AI system with data collected by another entity, it is necessary to distinguish:

- **the data diffuser:** the natural or legal person, public or private, who uploads online personal data or a dataset that contains personal data;
- **the re-user of the data:** the natural or legal person, public or private, who processes such data or datasets with the intention of using them on its own account.

The diffuser and the re-user of the data are, in principle, responsible for separate processing, since each determines the objectives and the essential means of its own processing.

The data diffuser is, in principle, responsible for the public dissemination, while the provider of the AI system that re-uses the data is responsible the usage of the data it has. The diffuser is not, in principle, responsible for the re-use of its data. It may, however, lay down conditions for the use of the data disseminated to limit its reuse or provide for certain provisions.

Example:

An administration makes real estate data publicly available and freely reusable (open data). A company wants to reuse this data to create a training dataset in order to develop an AI system able to predict certain real estate developments in a given area. The diffuser and the re-user are then responsible for separate processing, provided that these two processings are independent.

Find out more: Sheet 1 of the guide on the opening and reuse of publicly accessible data.

Joint controllers

The principle

When two or more controllers jointly determine the purposes and means of processing, they are joint controllers.

This qualification may be difficult in the presence of several stakeholders having an influence on the determination of the purposes and means of the processing. In particular, stakeholders need to determine whether they process the data for their own and distinct purposes or for a common purpose.

In practice

When the training dataset of an AI system is **fed by more than one controller** for a **jointly defined purpose,** the controllers may be qualified as **joint controllers.**

Examples:

Case 1: academic hospitals developing an AI system for the analysis of medical images choose to use the same federated learning protocol. The latter allows them to exploit data for which they are initially separate controllers, in order to benefit from the mutualization.
Together, they determine the purpose (training a medical imaging AI system) and the means of this processing (through the choice of the protocol and the determination of the data they exploit): they are therefore jointly responsible for this training processing.

Case 2: a consortium consisting of a municipality, a company providing automated image processing software and a company providing video devices is conducting an experiment to install enhanced cameras to record and analyse the flow and behaviour of vehicles using a traffic lane within the municipality. The contract between the city and the two companies provides for the use of the software by the municipality in real-time conditions and the possibility for companies to improve the automated image processing software by the data collected in real time. This improvement of the automated processing software benefits both the municipality and the companies providing automated image processing software and video devices.

The municipality and the two companies would thus be jointly responsible for the processing of the training dataset of the automated image processing software, provided that they jointly decide on the purpose and essential means of the processing and the companies do not act solely on behalf of the municipality. Indeed, it is possible to consider that they jointly decide on the essential means of processing (by choosing to feed the AI system training dataset with real-time data collected by enhanced cameras and data already collected by the company providing the automated image processing software) and the purpose of the processing (experimentally train an AI system that detects particular vehicle behaviour and improves the automated image processing software).

Conversely, if one of the companies intends to reuse the data for its own purpose, which it would be the only one to benefit from (e.g. in a research and development framework), then it could be considered that it is responsible for a separate processing.

In case of joint control, the parties must ensure the lawfulness of the processing (i.e. its compliance with the law), including by defining in a transparent manner their respective obligations under an agreement. The form of this agreement is not specified by the GDPR. The agreement must reflect the roles of each of the stakeholders, with joint controllers having to clearly specify "who does what" to ensure the protection of the data processed.

Please note: regardless of the terms of the agreement, the data subject may exercise his or her rights vis-à-vis each of the joint controllers.

The use of a processor

The principle

The processor is the natural or legal person who processes data on behalf of the controller, in the context of a service or provision.

In practice

The qualification of the AI system provider must be assessed on a case-by-case basis.

An AI system provider may be a processor when it develops an AI system on behalf of one of its customers as part of a service. The customer is the data controller as soon as they determine the purpose and means of the processing.

In other cases, the AI system provider may be the controller of the systems it designs to market them.

An AI system provider may use a provider to collect and process the data according to its documented instructions (e.g. to collect publicly available data on the Internet, reuse a specific dataset made available online, etc.). **The latter then qualifies as a processor**. It is essential for the provider of the AI system, as the controller, to ensure that its processor complies with the GDPR and limits the processing of data to its instructions, in particular by concluding a data processing agreement.

Moreover, the fact of using the same dataset for several customers, in the context of separate services, is generally a decisive indication that the provider is responsible for a separate processing, at least for the establishment of the database.

Example:

A provider has been entrusted with the creation of a training dataset by an AI system provider who has indicated precisely how it should be developed (in particular with regard to data sources and categories, with quality and documentation requirements). This service provider is likely to act as a processor.

Conversely, a service provider which, on its own initiative, would have created a dataset which it operates by developing AI systems adapted to the needs of each of its customers will likely be responsible for the processing of this dataset, regardless of its role in the specific processing carried out for those customers (which it could implement as a processor, for example on the basis of data provided by the customers themselves).

< Previous : Defining a purpose

Table of content

Next : Ensuring the lawfulness of the data processing - Define a legal basis >

Ensuring the lawfulness of the data processing - Defining a legal basis

07 juin 2024

An organisation that wishes to build a training dataset containing personal data and then use it to develop an AI system must ensure that the processing is lawful. The CNIL helps you determine your obligations based on your responsibility and the means of collecting or reusing the data.

The controller must in all cases define a legal basis and carry out, depending on the method of collection or re-use of the data, certain additional verifications.

There are several ways to build a training dataset, which can be used cumulatively:

- data is collected directly from individuals;
- data is collected from open sources on the Internet for this purpose;
- data was initially collected for another purpose by the controller itself (e.g. in the context of providing a service to its users) or by another controller. **This involves taking additional precautions.**

Define a legal basis

The principle

Like any personal data processing, the creation and use of a training dataset containing personal data can only be implemented if it corresponds to one of the “legal bases” provided for in the GDPR.

The legal basis is what gives an organisation the right to process personal data. The choice of the legal basis is therefore an essential first step to ensure compliance of the processing. Depending on the legal basis, the obligations of the organisation and the rights of individuals may vary.

The most relevant legal bases for training an algorithm are detailed below.

In practice

The determination of the legal basis must be carried out in a manner appropriate to the situation and the type of treatment. In order to establish a dataset for the training of an AI system, **the following legal bases may be envisaged in particular:**

The legal basis for consent

To be valid, the consent of the data subjects must meet four cumulative criteria: **it must be freely given, specific, informed and unambiguous**. The controller must be able to demonstrate the validity of the use of this legal basis by ensuring that each of these conditions, specifically defined by the GDPR, is met.

Example: an organisation wishes to film or photograph volunteers to create a dataset of images to train a system to detect certain specific gestures. It may base the processing on the basis of their consent.

When creating a dataset for, an organisation must ensure the **validity of the consent** collected.

Beyond the obligations of transparency, a certain amount of information must be provided to the data subjects before they consent, in order to enable them to make informed decisions and to allow them to withdraw their consent.

Consent must relate to a specific purpose (see how-to sheet 2 on the definition of the purpose).

The freedom of consent implies, in principle, the possibility for data subjects to give their consent in a granular way, where there are different purposes.

Example: the consent of individuals to the use of their image, collected at a company event for communication purposes, does not mean that they consent to a re-use of the data for building a training dataset or improving an AI system. In this case, two separate consents must be collected (e.g. via two check boxes).

The freedom of consent may also be impacted in the case of an imbalance of power in the relationship between the data subject and the controller, especially if the controller is a public authority or an employer.

Example: a company wants to use the data of its employees to develop an AI system. Their consent can only be validly collected in exceptional situations, where they are able to refuse to give their consent without fear or incurring negative consequences. As controller, the company must ensure, in any event, that the communications intended to present the device to employees are neither incentive nor binding. It must inform the volunteers of the possibility of no longer participating in the collection of their data at any time, without any consequence.

It does not seem possible to obtain valid consent in some cases. This is often the case when the controller collects data accessible online or reuses an dataset available online, especially given the lack of contact with the data subjects and the difficulty in identifying them. In these cases, the controller must rely on a more appropriate legal basis.

There may also be difficulties related to the right to withdraw consent, for example due to technical obstacles to the identification of data subjects. If it is not possible for the controller to guarantee the possibility of exercising this right, it is recommended to rely on another legal basis.

The legal basis for the legitimate interest

The controller may rely on its legitimate interest provided that it complies with the following conditions:

- **the legitimacy of the interest pursued by the controller.** For example, the interest of an organisation in developing a model for the commercialisation of an AI system or in order to contribute to the improvement of scientific knowledge, for example by publishing the tools developed (code, model, experimentation protocol, etc.) and research results.

- **the necessity of the data processing.** For example, processing for the purpose of creating up a training dataset containing images of people may be considered necessary for the interests of an organisation wishing to develop a pose estimation system, where anonymous or synthetic data are not sufficient.

- **the absence of a disproportionate impact on data subjects' interests and rights and freedoms,** taking into account their reasonable expectations. Balancing of the rights and interests at hand depends on the specific characteristics of the processing and in particular on the safeguards implemented to ensure the best possible balance between those interests and to limit the impact of the processing on the data subjects.

More often than not, creating a training dataset whose use is lawful can be considered legitimate. However, an analysis is necessary to determine whether the use of personal data for this purpose does not disproportionately infringe the privacy of the data subjects, even when the data is not nominative. To guarantee that its processing is proportionate, the controller may implement measures such as: pseudonymisation of the data, ensuring the absence of sensitive data, defining selection criteria to limit the collection to the relevant and necessary data, etc.

Examples:

A company wants to develop an AI system that can predict a person's psychological profile from online data that may relate to them. Its commercial interest in developing such a system is likely to be insufficient in the light of the interests, rights and freedoms of data subjects: another legal basis will have to be sought or the project abandoned.

An organisation creates a training dataset by collecting comments made public and freely accessible by online users on forums, blogs and websites. The purpose of this processing is to design an AI system to assess and predict the appreciation of works of art by the general public. In this case, its interest in developing and possibly marketing an AI system may be considered legitimate. The collection of feedback on the works may be considered necessary for the development of the model, especially given the amount of training data required . It should be noted that the legal basis of legitimate interest gives data subjects the right to object to the processing of their data (for reasons relating to their particular situation).

The legal basis of the task carried out in the public interest

The possibility of relying on the legal basis of the "task carried out in the public interest" implies:

- that the task of processing is provided for in a normative text applicable to the controller;
- that the use of the data makes it possible to carry out this task specifically, in a relevant and appropriate manner.

Examples:
Researchers from a public research laboratory on the French language wish to analyse the evolution of the use of the language online. To do so, they create a dataset based on comments published online on different social networks (anonymised at short notice) in order to train a model that automatically detects and analyses the occurrence of certain expressions or spelling forms.

To the extent that the controller is a public laboratory, in this case the researchers may base the data processing on the task carried out in the public interest. This legal basis can be used, in general, for data processing carried out by public or private research laboratories entrusted with a task of public interest, when the processing is necessary for their research activity.

The Pôle d'Expertise de la Régulation Numérique (PEReN) is authorised to reuse, under certain conditions, publicly accessible data from certain platforms in order to carry out experiments aimed in particular at designing technical tools for the regulation of online platform operators, in accordance with Article 36 of Law No 2021-1382 of 25 October 2021 and Decree No 2022-603 of 21 April 2022.

For more information:

- Use case sheet 4 of the guide on the re-use of publicly accessible data (open data)
- What legal basis for research processing?

The legal basis of the contract

The legal basis of the contract could be used for the creation of a training dataset for an AI system provided that a valid contract is concluded between the controller and the data subject and that the processing is objectively necessary for its performance.

Contracts concluded for this purpose must comply with other applicable rules, such as labour law or intellectual property.

Examples:

A text editor software company offers an automated and personalised mail generation service, to which the user contractually subscribes, and for which the editor collects the data of the users of this service. The data processing for this personalisation service may be considered, subject to its specific characteristics, necessary for the performance of the contract.

Conversely, the operator of an online social network registered in its general terms and conditions of use that it intends to reuse the data of its users (provided by them, observed or inferred by the operator) to develop and improve new products, services and functionalities useful for its users. It cannot base the processing on the legal basis of the contract since such processing is not objectively necessary in order to offer them its online social network service (ECJ, 4 July 2023, Meta Platforms Inc. and a. c/Bundeskartellamt, C-252/21).

Sensitive data: prohibited processing, with exceptions

Sensitive data is a particular category of personal data defined in Article 9 of the GDPR. Sensitive data includes, for instance, data revealing the alleged racial or ethnic origin of the data subjects, or biometric data for the purpose of uniquely identifying a natural person, such as a facial template.

The GDPR prohibits the processing of such data, except in the cases listed in Article 9.2. of the GDPR. These exceptions include in particular:

- processing operations for which data subjects gave their explicit consent (active, explicit and preferably written, freely given, specific and informed);

- processing of personal data which is manifestly made public by the data subject;

  In its Guidelines on targeting users of social networks, the EDPB provides a list of factors to be taken into account in determining whether the data is manifestly made public: the default setting of the social media platform, the nature of the platform, the accessibility of the page concerned, the visibility of the information about its public nature, whether the data subject has published the data himself or whether it has been published by a third party or deduced.

  It is important to check whether the data subject wished, explicitly and by a clear positive act, on the basis of an informed setting, to make his or her personal data accessible to the general public or, on the contrary, to a more or less limited number of selected persons (ECJ, 4 July 2023, Meta Platforms, C-252/21).
- processing necessary for reasons of substantial public interest, on the basis of EU or Member State law;

- processing operations necessary for the purpose of scientific research on the basis of European Union or Member State law.

Particular attention should be paid to the collection of sensitive data when using web scraping tools that involve the processing of large volumes of data. The controller has to implement measures to automatically exclude the collection of irrelevant sensitive data, in particular by applying filters to exclude the collection of certain categories of data or to exclude certain sites that gather sensitive data by nature. If, despite the measures taken, the organisation processes incidentally and residually sensitive data that it had not sought to collect, it is not considered illegal. In particular, the Court of Justice of the European Union held that that prohibition applies to the operator of a search engine "in the context of its responsibilities, powers and possibilities" (ECJ, Grand Chamber, 24 September 2019, GC and Others, C-136/17). On the other hand, if the organisation comes to know that it is processing sensitive data, it has to proceed, as far as possible, to its immediate and automated deletion.

Please note:

- A how-to sheet on bias management will be published at a later date. It will clarify the possibility of processing sensitive data for the purpose of detecting and correcting bias in the training dataset.
- The CNIL is currently conducting work on the issue of AI in health, which will be published later.

The basis of the legal obligation

This legal basis may seem relevant in some cases for data processing carried out in the deployment phase, since an AI system can sometimes be used by the controller to comply with a legal obligation (provided that it requires the processing of personal data).It is however more difficult to rely on this basis for its development.

Indeed, in order to rely on this legal basis, the processing must be necessary to meet a specific legal obligation to which the controller is subject. The text on which it is based must at least define the purpose of the processing and may frame it more precisely (in particular through the types of data to be processed, the limitation of the purposes or other conditions to be respected). The more precise the legal obligation, the easier it is to justify why it requires the processing of personal data.

However, since legal obligations are generally not specific enough to provide for the development of AI systems, it will most often be necessary to rely on another legal basis to develop such systems.

Examples:

In the insurance sector, actuarial studies are based on mathematical, probabilistic and statistical models similar to AI systems, the objective of which is to identify, qualify and quantify risks (and associated amounts) related to insurance contracts.

Since the general solvency obligations of insurance institutions are not sufficiently precise, it is not possible to consider that the development of such systems is necessary for their compliance. The legitimate interest then appears to be the most relevant legal basis.

< Previous : Determining the legal qualification of AI system providers

Table of content

Next : Ensuring the lawfulness of the data processing - In case of re-use of data >

Ensuring the lawfulness of the data processing - In case of re-use of data, carrying out the necessary additional tests and verifications

07 juin 2024

If it re-uses data previously collected, the data controller is required to carry out certain additional verifications to ensure that that the processing is lawful. The CNIL helps you determine your obligations, depending on the means of collecting the data and its source.

The principle

In some cases, depending on the methods of collection and the source of the data used for the creation of the training dataset, the controller is required to carry out certain verifications to ensure that the processing of data is lawful. These verifications must be carried out in addition to the identification of the legal basis for the processing.

In practice

The provider reuses the data it originally collected for another purpose

A data controller may wish to reuse the data it has collected for an initial purpose (e.g. in the context of providing a service to individuals) in order to create a training dataset.

In that case, **it must determine whether that further processing is compatible with the purpose for which the data were originally collected**, where the processing is not based on the data subject's consent or on Union or Member State law.

The obligation to carry out this "compatibility test" applies to the further processing of data (within the meaning of Article 6.4 GDPR), i.e.:

- which have not been foreseen or brought to the attention of data subjects when collecting the data;
- which are carried out by the same controller who decides to re-use data for a purpose distinct from the purpose for which it was collected, including when it comes to publishing it on the Internet or sharing

it with third parties for re-use for another purpose.

Please note: no compatibility test is required for the intended purposes and brought to the attention of the data subjects as soon as they are collected in accordance with the principle of transparency, including where some of them may appear secondary or accessory. For example, the sharing of data by a controller with its processor for the improvement of the performance of its algorithm does not require a compatibility test, if this purpose was intended and brought to the attention of the data subject (subject to its compliance with the conditions of legality for this purpose of improving the algorithm).

In order to carry out this “compatibility test”, it must take into account in particular:

- the existence of a link between the initial purpose and the purpose of the intended further processing;
- the context in which the personal data was collected, in particular the reasonable expectations of the data subjects, depending on the relationship between the data subjects and the controller;
- the type and nature of the data, in particular according to its sensitivity (biometric data, geolocation data, concerning minors, etc.);
- the possible consequences of the envisaged further processing for the data subjects;
- the existence of appropriate safeguards (such as encryption or pseudonymisation).

Examples:

The provider of a text editor launches a generative AI feature to complete certain sentences or paragraphs. Some time after the deployment of this feature, it wishes to re-use the manual corrections made by users to the content of the texts thus generated, in order to offer each user a personalised version of their recommendation service (for example to better understand and anticipate the way they write) on the basis of their respective data.

A video streaming platform is now considering reusing the history and playlists it has saved as part of the provision of the service to offer each user a personalised version of their referral service (e.g. to better anticipate and understand their preferences) based on their respective data.

In both cases, the new purpose may be considered compatible with the original purpose of the provision of the service, provided that the guarantees implemented are sufficient (e.g. through the possibility of objecting such re-use without having to provide any grounds) on the basis of their respective data.

Where the reuse of the data pursues statistical or scientific research purposes, the processing is presumed to be compatible with the original purpose if it complies with the GDPR and if it is not used to make decisions regarding the data subjects. The compatibility test is therefore not necessary.

In order to pursue a statistical purpose within the meaning of the GDPR, the processing must only aim at the production of aggregated data for themselves: the sole purpose of the processing must be the calculation of the data, their display or publication, their possible sharing or communication (and not taking subsequent decisions, individual or collective). The statistical results thus obtained must constitute aggregated and anonymous data within the meaning of the data protection regulations. The use of statistical techniques of machine learning is not enough to consider that they are processing “for statistical purposes”, since the purpose of processing is not to produce aggregated data for themselves. The use of these techniques is more of a mean to train the model.

The notion of “scientific research” is broadly understood in the GDPR. In summary, the aim of the research is to produce new knowledge in all areas in which the scientific method is applicable. Any processing of data for scientific research purposes must be subject to appropriate safeguards for the rights and freedoms of the data subject, such as anonymisation or pseudonymisation (referred to in Article 89 GDPR).

Learn more:

- Scientific research (excluding health) [In French]

Please note: even when the further processing is compatible, a valid legal basis must always be identified and the data subjects informed, in particular in order to be able to exercise their rights.

Focus: under what conditions can a dataset originally constituted for scientific research purposes be reused?

The GDPR facilitates the reuse of data for scientific research purposes: this reuse is considered compatible with the original purpose of the processing and certain derogations (in particular to the rights of individuals) are possible.

On the other hand, where a controller has processed data for scientific research purposes and intends to reuse them for other purposes (on its own behalf or to transmit it to a third party), it must comply with certain conditions.

- **The reuse of a dataset will be possible:**
  - **if the data have been previously anonymised, or**
  - **if the reuse is compatible with the purpose for which the controller collected the data** (according to the “compatibility test” detailed above) **and the new processing is implemented in compliance with the GDPR** (information to individuals about this new purpose, identification of a legal basis, etc.). The derogations allowed by the GDPR for scientific research will no longer be mobilised.

    In the event of transmission of the data to third parties, the compatibility of further reuse with the research purpose may be guaranteed in particular by a license.

The provider reuses a publicly accessible dataset

Datasets containing personal data may be freely made available on the Internet outside the legal framework for open data. Most often, it corresponds to data that were already publicly accessible and that constitute a dataset or corpus disseminated on the website of a university or a platform dedicated to sharing datasets, to facilitate their re-use.

Checking the lawfulness of making the dataset available online is primarily the responsibility of the controller who put the dataset online (where appropriate, by ensuring that it is a compatible further processing if it had not initially collected the data for that purpose). However, in order to be able to rely on a valid legal basis under the GDPR, the controller who reuses the data must ensure that they are not reusing a dataset whose creation was manifestly unlawful (e.g. from a data leak).

The re-user may not re-use an established or uploaded dataset for which they cannot ignore that they do not comply with the GDPR (Article 5.1.a GDPR) or other rules, such as those prohibiting breaches of the security of information systems or infringements of intellectual property rights.

In addition, the person who downloads or reuses a manifestly illegal dataset may be guilty of the offence of concealment (Article 321-1 of the French Criminal Code).

The possibility of reusing freely a dataset made available on the Internet is not necessarily subject to in-depth verifications on compliance with all GDPR rules or other applicable legal rules (copyright, data covered by business secrecy, etc.), which are primarily under the responsibility of the organisation that uploads the data. However, an organisation cannot reuse a dataset that would be manifestly unlawful.

This obvious illegality must be assessed on a case-by-case basis. As such, the CNIL recommends that re-users ensure that:

- **The description of the dataset mentions their source.**

Example: a dataset whose description would explain that it was made from publications on a professional social network referred to by name.

Conversely, if a dataset containing video surveillance images does not specify its sources, such a dataset should not be reused until further details have been obtained to remove doubts as to the conformity of its constitution and dissemination;

- The creation or dissemination of the dataset **is not manifestly the result of a crime or an offence nor has been the subject of a public conviction or sanction** by a competent authority which involved the deletion or prohibition of further use of the data;

Examples: a company wishes to create a training dataset to develop a recommendation AI system for its consumers. If it acquires for this purpose a dataset on the dark web from, for example, an infringement of an automated processing system punishable by law (within the meaning of Article 323-1 of the French Criminal Code), it cannot ignore its criminal origin. In this case, the illegality of the dataset would then be obvious.

The same applies to a company wishing to reuse a dataset for which a court decision has found an infringement of an intellectual property right such as that of dataset producers (within the meaning of Article L. 342-1 of the French Intellectual Property Code);

- there is no clear **doubt that the dataset is lawful (in particular that the initial processing is not manifestly lacking a legal basis** when the data are so intrusive that they cannot be processed without the consent of the individuals), ensuring in particular that the conditions for collecting the data are documented enough;

Examples: on a hosting platform for ML practitioners, a company identifies a dataset containing the home-to-work journeys of thousands of people. Its description explains that it is accurate geolocation data, not anonymous, without detailing the source. In that case, the company cannot ignore that there is a serious doubt as to the lawfulness of the dissemination of such a dataset without the consent of the data subjects.

Conversely, it would be possible to create a dataset from another dataset whose description leaves no clear doubt as to its lawfulness. For example, a pseudonymised dataset, initially made public by data subjects on an identified website and that does not contain sensitive data.

The same applies to the reuse of an aggregated dataset that the diffuser would present as anonymous. For example, an organisation that wishes to create a training dataset to develop an AI system able to predict the socio-economic impact of population ageing could reuse anonymous aggregated datasets containing demographic information (number of active persons, age of persons, fertility rate or elderly dependency rate).

- the dataset does **not contain sensitive data** (e.g. health data or political opinions) or **infringement data** (as defined in Articles 9 and 10 GDPR) or, if it contains such data, it is recommended to carry out additional verifications to ensure that such processing is lawful (mainly for sensitive data to ensure explicit consent of data subjects, or that the data have been clearly made public by the data subjects as specified below and for data relating to infringements that such use is made possible by the French Data Protection Act).

Example: on an online forum, a researcher discovers a non-anonymous dataset that would contain, according to his description, the healthcare information of a hundred patients with a particular pathology and who would come from French hospitals. In this case, the researcher should seriously doubt whether the dissemination of this dataset is lawful in view of the supervision of health data provided for by the GDPR and the French Data Protection Act.

Such prior verifications could be included in the Data Protection Impact Assessment (DPIA).

Certain failures committed by the controller to create and disseminate a dataset do not systematically and irreparably impact the lawfulness of the processing carried out by the re-user. Thus, **a re-user may use a dataset with minor illegalities**, provided that the reuse meets the requirements of the GDPR.

Example: the provision of incomplete information when creating or disseminating the dataset, or a lack of adequate documentation of the compliance of these processing operations (which it is necessary to verify with the diffuser or publisher of the dataset).

The provider reuses a dataset acquired from a third party (data brokers, etc.)

Some providers wish to create a training dataset from datasets owned by third parties.

For the third party who shares personal data, this means ensuring the lawfulness of this transmission

- **Case 1: the data was collected specifically to be shared to create a training dataset**

The third party **will have to ensure that the processing of data transmission complies with the GDPR** (definition of an explicit and legitimate purpose, identification of a valid legal basis, information to data subjects and management of the exercise of their rights, etc.) for which they assume responsibility.

- **Case 2: the third party did not initially collect the data for this purpose**

Where the third party initially collected the data for other purposes (e.g. in the context of the provision of a service to data subjects), it **is for the third party to ensure that the transmission of such data is for a purpose compatible with the purpose(s) which justified its collection.** It will therefore have to carry out a “compatibility test”.

Note that the initial owner of a dataset sometimes authorises its use under a license agreement that provides for its terms and conditions (in particular under intellectual property law). This license agreement can, for example, regulate this compatibility by limiting possible reuse.

For the re-user, this usually involves a series of verifications of the initial controller’s processing

The controller **must ensure that they do not re-use a dataset whose creation or transmission was manifestly unlawful** (for example, in the absence of an indication as to its source, in case of blatant doubt as to its lawfulness, in particular in the case of the processing of sensitive data, etc.). This results from the general principle of lawfulness of processing in Article 5.1(a) GDPR, in addition to the risk of being guilty of the offence of concealment (Article 321-1 of the French Criminal Code). This implies for the controller to carry out at least the same verifications as those set out in the section above.

The re-user of a dataset transmitted by a third party will be all the less likely to ignore that the dataset was created or shared in breach of the GDPR or of more general rules (such as those prohibiting breaches of the security of information systems or infringements of intellectual property rights) since its relationship with that third party allows it to remove any doubts that it may have.

An agreement between the initial data holder and the re-user is thus recommended in order to enable the latter to ensure the lawfulness of its own processing, even if it is not explicitly required by the GDPR.

In this regard, the CNIL recommends providing a number of indications in the contract such as:

- the source, the context of the data collection, the legal basis for the processing and the data protection impact assessment (see in particular how-to sheet 5 on the implementation of a DPIA) if necessary, in order to avoid the risks of having an unlawful dataset;
- the information provided to the data subjects (especially with regards to the purpose and the recipients);
- any guarantees as to the lawfulness of this data sharing by the original data holder (e.g.: the compatibility of the purpose, the lawfulness of sharing, etc.).

The CNIL provides a description sheet of the dataset that can be used for this purpose.

Please note: if the re-user wishes to base his processing on consent obtained by a third party, they must be able to prove that valid consent has indeed been obtained from the data subjects. The obligation to provide proof of consent cannot be fulfilled by the mere presence of a contractual clause requiring one of the parties to obtain valid consent on behalf of the other party. Such a clause does not allow the organisation to guarantee, in all circumstances, the existence of valid consent (see the CNIL's deliberation no. SAN-2023-009 of 15 June 2023). The contract may, on the other hand, be used to frame:

- the mechanisms put in place to demonstrate the collection of valid consent;
- the provision of evidence to the organisation wishing to rely on the consent of data subjects;
- where applicable, the conditions under which such evidence must be retained, in particular in order to maintain its probative value.

Example: the provider of a generative AI image system approaches a data broker to create a training dataset including photographs.

To this end, they enter into a contract that guarantees the provider the lawfulness of the data shared and regulates the provision of crucial indications for the compliance of its processing (e.g. evidence of the context of the collection of data in order to assess its legitimate interest, guarantees in relation to other regulations such as governing the assignment of intellectual property rights, etc.).

In addition to these prior verifications, and regardless of the method of collection used, **re-users must ensure that their own processing are fully compliant.**

It should be noted that this obligation also applies when they reuse datasets whose creation and dissemination do not fall within the scope of French or European law. For more information on the territorial scope of the GDPR, see the how-to sheet "What is the scope of the AI how-to sheets".

In particular, the re-user must ensure compliance with the requirements regarding the persons whose data is present in the dataset thus obtained. They must inform them of the processing that they wish to implement, and allow them to exercise their rights.

Please note: a dedicated sheet on the reuse of personal data will be published later. It will complement the elements introduced in this sheet, in particular with practical case studies.

< Previous : Ensuring the lawfulness of the data processing - Defining a legal basis

Table of content

Next : Carrying out a data protection impact assessment if necessary >

Carrying out an data protection impact assessment if necessary

07 juin 2024

Creating a dataset for the training of an AI system can create a high risk to people's rights and freedoms. In this case, a data protection impact assessment is mandatory. The CNIL explains how and in which cases it should be realised.

The Data Protection Impact Assessment (PDIA) is an approach that allows to map and assess the risks of a personal data processing and to establish an action plan to reduce them to an acceptable level. This approach, facilitated by the tools provided by the CNIL, is particularly useful to control the risks associated with a processing before it is implemented, but also to ensure their follow-up over time.

In particular, a DPIA makes it possible to carry out:

- an identification and assessment of the risks for individuals whose data could be collected, by means of an analysis of their likelihood and severity;
- an analysis of the measures enabling individuals to exercise their rights;
- an assessment of people's control over their data;
- an assessment of the transparency of the data processing for individuals (consent, information, etc.).

The DPIA must be carried out prior to the implementation of the processing and should be changed iteratively as the characteristics of the processing and risk assessment evolve.

The realisation of a DPIA for the development of AI systems

Identifying when a DPIA is needed

The development of AI systems requires, in some cases, the realisation of a DPIA if the envisaged processing is likely to create a high risk to the rights and freedoms of natural persons (Article 35 GDPR).

In its guidelines on the DPIA, the European Data Protection Board (EDPB) has identified nine criteria to assist data controllers, i.e. the AI system providers, in determining whether a DPIA is required. Any processing of personal data fulfilling at least two criteria on this list should be presumed to be subject to the obligation to carry out a DPIA. Some of these criteria are particularly relevant for processing taking place during the development phase:

- the collection of sensitive data or highly personal data (e.g. categories of data that can be considered to increase the risk of harm to the rights and freedoms of individuals, such as location data or financial data);
- the large-scale collection of personal data;
- the collection of data from vulnerable persons, such as minors;
- the crossing or combination of data sets;
- innovative uses or application of new technological or organisational solutions.

In all cases, it is necessary to consider the existence of risks for persons as a result of the establishment of a training dataset and its use: if there are significant risks, in particular due to data misuse, data breach, or where the processing may give rise to discrimination, a DPIA must be carried out even if two of those criteria are not met; conversely, a DPIA does not have to be carried out if several criteria are met but the controller can establish with sufficient certainty that the processing of the personal data in question does not expose individuals to high risks.

On the basis of these criteria, the CNIL has published a list of personal data processing for which the realisation of a DPIA is mandatory (for more information, see the CNIL’s website). Of these, several may rely on artificial intelligence systems, such as those involving profiling or automated decision-making: in this case, a DPIA is always required.

Is the use of an artificial intelligence system an “innovative use”?

Innovative use is one of the 9 criteria that can lead to the realisation of a DPIA: it is assessed in the light of the state of technological knowledge and not only of the context of the processing (a processing can be very “innovative” for a given organism, because of the technological novelty it brings to it, without, however, being an innovative use in general). The use of artificial intelligence systems **is not systematically a matter of innovative use or the application of new technological or organisational solutions.** All processing using an AI system will therefore not meet this criterion. In order to determine whether the technique used falls within such uses, it is necessary to distinguish between two categories of systems:

- Systems that use AI techniques that have been experimentally validated for several years and tested in real-life conditions. These systems are not part of the innovative use or application of new technological or organisational solutions.

Example: certain regression or clustering techniques or model architectures such as random forests, in cases where the risks associated with their use are known;

- Systems that use new techniques, such as deep learning, and whose risks are just beginning to be identified today, but are still poorly understood or mastered. These systems are part of innovative use.

Example: generative AI systems trained on large amounts of data whose behaviour cannot be anticipated in all situations.

By way of illustration, a research project aimed at developing automatic language processing tools for clinical applications in the medical field, based on large volumes of data (transcript of audio data, clinical studies, medical results, etc.), can be an innovative use, especially given the uncertainty as to the results to be obtained.

Is the training of an artificial intelligence system a "large-scale" processing?

Large-scale collection is one of the 9 criteria that can lead to the implementation of a DPIA: while the development of an AI system often relies on the processing of a large amount of data, this does not necessarily fall within the scope of large-scale processing which aims to "*process a considerable amount of personal data at regional, national or supranational level [and which may] affect a significant number of data subjects*" (recital 91 GDPR). For AI systems, in particular, it will be necessary to determine whether the development involves a very large number of people.

Examples:
A research organisation wants to build a large dataset of landscape photos (mountain, ocean, desert, cities, etc.) to improve the performance of computer vision systems. Some of these images feature images of individuals, sometimes recognizable.

Even if the dataset has millions of images covering the entire surface of the planet, if the number of images containing recognizable individuals (and therefore personal data) is limited (for example to a few thousand), the processing will not be called "large-scale processing". However, it is not excluded that a DPIA may be required according to the other criteria to be verified.

Where a provider of a conversational agent constitutes a dataset to train its language model (LLM) from a considerable volume of publicly accessible personal data on the Internet collected through web scraping techniques, the processing can be described as "large-scale processing".

Risk criteria introduced by the EU AI Act

The European AI Act aims to provide a legal framework for the development and deployment of AI systems within the European Union. It distinguishes several categories of systems according to their level of risk: prohibited systems, high-risk systems, systems requiring transparency guarantees and minimum risk systems. **The CNIL considers that for the development of all the high-risk systems covered by the AI Act, the realization of a DPIA will be presumed necessary when their development or deployment involves the processing of personal data.**

The realization of the DPIA may be based on the documentation required by the AI Act provided that the elements required by the GDPR (Article 35 GDPR) are included. The elaboration of more precise rules on the relationship between these requirements is the subject of European work in which the CNIL actively participates and which will be the subject of subsequent publications. This work will aim, in particular, to avoid any duplication of obligations on actors by prioritising the reuse of the elements constituted from one framework to another.

Moreover, **the CNIL considers that the development of a foundation model or a general-purpose AI system, in that their uses cannot be exhaustively identified in the majority of cases requires the realisation of a DPIA when it involves the processing of personal data.** Indeed, although these models and systems are not considered to be high risk by default by the AI Act, their dissemination and their future uses could entail risks for those whose data were processed during development, or for the persons concerned by their use.

The realization of a DPIA for foundation models and general purpose AI systems will facilitate the compliance of the processing implemented by their users. In this respect, the sharing or publication of the realized DPIA may facilitate the compliance of all the actors involved, in particular in the case of the dissemination of open source models, or the provision of systems for all.

Defining the scope of the DPIA

The scope of the DPIA may differ depending on the provider's knowledge of the use that will be made, by itself or by a third party, of the AI system it develops.

Where the operational use of the AI system in the deployment phase is identified from the development phase

If the system provider is also the data controller for the deployment phase and if the operational use of the AI system in the deployment phase is identified from the development stage, it is recommended to carry out a general DPIA for the entire processing. The supplier will then be able to supplement this DPIA with the risks associated with both phases.

If the provider is not the data controller for the deployment phase but identifies the purpose of use in the deployment phase, it may propose a model of DPIA accordingly. This may allow it, in particular, to take into account certain risks that are easier to identify during the development phase. However, the user of the AI system, as controller, remains obliged to perform a DPIA, for example on the basis of the provider's template.

It should be noted that, in some cases, it is not possible to determine precisely and in advance the supervision of the deployment phase. For example, some risks can be reassessed after a calibration phase of the AI system under its deployment conditions. **The DPIA will then have to be modified iteratively** as the characteristics of the processing are defined at the deployment stage.

Where the operational use of the AI system in the deployment phase is not clearly identified in the development phase

In this case, the provider of the system will only be able to carry out its impact assessment on the development phase. It will then be up to the controller of the deployment phase to analyse, with regard to the characteristics of the processing, whether a DPIA is necessary for that phase. If the purposes of the deployment phase are multiple, the controller may decline the same general DPIA for each of the specific use cases

AI Risks to consider in a DPIA

Processing of personal data based on AI systems present specific risks that should be taken into account:

- the risks to data subjects related to misuse of the data contained in the training dataset, in particular in the event of a data breach;

- the risk of automated discrimination caused by the AI system introduced during development, for example linked to a lower performance of the system for certain categories of people;

- the risk of producing fictitious content on a real persons, which is particularly important in the case of generative AI systems, and may have consequences for their reputation;

- the risk of automated decision-making caused by automation or confirmation. This risk may arise in particular if the necessary explanatory measures are not taken during the development of the solution (such as the use of a trust score, or intermediate information such as saliency map), thus limiting the ability of the agent using the system to verify its performance under real conditions. This risk may also arise when the staff member is unable to take a decision contrary to the outputs of the system without prejudice to them (due to hierarchical pressure, for example);

- the risks associated with known attacks specific to AI systems such as attacks by data poisoning, by inserting a backdoor, or by model inversion;

- the risks related to the confidentiality of the data that could be extracted from the AI system;

- Systemic and serious ethical risks related to the deployment of the system, such as impacts on the democratic functioning of society, or respect of fundamental rights (e.g. in cases of discrimination), which can be taken into account during the development phase.

- Finally, the risk of users losing control over their published online and freely accessible data, as large-scale collection is often necessary for training an AI system, in particular when it is collected by web scraping;

When several data sources are used for the development of the AI system, the risks mentioned here are to be taken into account for each source, but also for the overall set thus constituted. Moreover, where the system is developed on the basis of a pre-trained model provided by a third party, the model must still be subject to the risk analysis described above, for example on the basis of information provided by the body providing the model.

Finally, analyses from benchmarks published by the CNIL or by third parties may be integrated or associated with the DPIA. Among these benchmarks, the CNIL recommends using:

- the self-assessment guide published by the CNIL;

- the benchmarks and frameworks identified by the CNIL on the page “Other guides, tools and good practices ”;

- the EU AI Act, and in particular its Annex IV detailing the technical documentation to accompany the placing on the market of high-risk AI systems.

Link between the documentation requirements of the AI Act and the implementation of a DPIA

While both documents are part of a risk anticipation logic and can overlap, there are significant differences between the DPIA and the documentation of compliance of the proposed AI Regulation.

On the one hand, they differ in their scope. Since some AI systems that are not classified as high-risk will rely on processing operations that pose risks to the protection of personal data, these will require the implementation of a DPIA.

On the other hand, it will be up to the controller in question, whether the latter concerns the development or deployment of the system, to carry out a DPIA, whereas the documentation requirements of the AI Act will essentially weigh on the provider of the AI system.

However, it is foreseen that in cases where an AI system provider subject to the documentation requirements of the AI Regulation is also required to carry out a DPIA, it is encouraged to include elements from the first document in the second document. The elaboration of more precise rules on the relationship between these requirements is the subject of European work in which the CNIL actively participates and which will be the subject of subsequent publications. They will explore the possibility of working only on a single document incorporating the requirements of the DPIA and AI Act documentation.

Find out more: The CNIL DPIA Guides

Actions to be taken on the basis of the results of the DPIA

The DPIA is an exercise that first determines the level of risk associated with the processing of personal data. Based on this level, a set of measures should be devised in the DPIA to reduce and maintain it at an acceptable level. These measures must incorporate the CNIL's recommendations that apply, whether they relate to AI techniques used or not.

Learn more: DPIA – Knowledge Bases

In addition, certain measures specific to the AI field – in particular of a technical nature – may be implemented, including:

- security measures, such as homomorphic encryption or the use of a trusted execution environment;
- minimisation measures, such as the use of synthetic data;
- anonymisation or pseudonymisation measures, such as differential privacy;
- data protection measures such as federated learning ;
- measures to facilitate the exercise of rights or remedies for persons, such as machine unlearning techniques, or measures to explain and trace outputs from AI systems;
- audit and validation measures, for example based on fictitious red teaming attacks, in particular to identify and correct biases or errors against certain persons or categories of persons.

Other, more generic measures may also be applied:

- organisational measures, such as supervision and limitation of access to training datasets, which may allow for a modification of the AI system, the limitation of access to data by third parties and subcontractors;
- governance measures, such as the setup of an ethical committee;
- measures of traceability of actions carried out in order to identify and explain abnormal behaviour;
- measures providing for internal documentation, such as the drafting of an IT charter.

These measures should be selected on a case-by-case basis in order to reduce the risks specific to the processing of data in question. They will need to be integrated into an action plan and monitored. In addition, being intended to protect data during the development of the AI system and in particular when setting up the dataset, they may be complemented by other AI-specific measures to be applied during the deployment phase. In particular, a description of the specific measures for the deployment of generative AI will be provided at a later stage.

Finally, the publication of all or part of the DPIA is recommended to improve transparency: while some parts of the DPIA do not have to be published to the extent that they may be covered by business secrecy or give confidential information on the security of the system, others present the risks and measures taken to limit them and their publication is of interest to system users and the general public.

< Previous : Ensuring the lawfulness of the data processing - In case of re-use of data

Table of content

Next : Taking into account data protection when designing the system >

Taking into account data protection when designing the system

07 juin 2024

To ensure that the development of an AI system respects data protection, it is necessary to carry out a prior reflection when designing it. This sheet details the steps involved.

When considering the design choices of an AI system, the principles of data protection, and in particular the minimisation principle, must be respected. This approach takes place at four levels. A controller must therefore ask itself about:

- **the objective of the system** it wishes to develop;
- **the method to be used** which will affect the characteristics of the dataset;
- **the data sources mobilized** (see the how-to sheet on the compliance of the processing with the law, on open sources, on third parties, etc.) and among these sources, **the selection of data strictly necessary**, in view of the usefulness of the data and the potential impact their collection has on the rights and freedoms of data subjects;
- **the validity of the choices** previously made. Such validation may take different (non-exclusive) forms, such as a **pilot study** or the solicitation of an ethics committee.

The objective of the system

The aim of this step is to design, on the basis of the identified purpose (see how-to sheet 2 Defining a purpose), a system that complies with a specification, while limiting the potential consequences for the data subjects.

By specifying the use of the AI system in the deployment phase (whether implemented directly by the provider or by a third party), the system provider must determine:

- the type of expected results/outputs;
- acceptable performance indicators of the solution, whether quantitative (F1-score, mean squared error, computation time) or qualitative indicators (e.g. from human feedback);
- the context in which the system is used in order to identify priority information for its operational use;
- excluded contexts of use and information not relevant to the envisaged main use case(s) of the system.

Some AI techniques can allow for complex tasks that go beyond the initial objectives of the provider. By precisely defining the expected functionality, it is possible to avoid risks of over-collection.

Example: for the purpose of training an AI system using video protection camera images to estimate the number of persons standing in a tram , the following systems are technically feasible:

- a neural network used for the detection of the presence of people in a wagon, without posture analysis, integrated into an algorithm making a count of standing people (the number of standing people can be inferred from the number of seating positions);

- a neural network performing an analysis of the posture of people in a wagon integrated into an algorithm counting standing people.

The first network could provide less information (including the count of standing persons). However, if the estimate given is sufficient for the intended use case, in particular for the calculation of occupancy statistics, then it is preferable to use this model. Indeed, it will require a smaller amount of data for its training and at the same time meet the objective pursued, while the latter requires the collection and annotation of more specific and larger datasets. The principle of minimisation then converges with the reduction of system design costs, without prejudice to the accuracy of the system.

The method to be used

Quite often, the same task can be carried out by different techniques. However, not all of them are equivalent, as they do not necessarily involve using the same data and the same amount. They may not make it possible to achieve the same level of performance, present more or less important challenges in terms of explainability, be subject to different operational constraints (such as the computational cost). While taking these issues into account, **the system provider must select the technique most respectful of the rights and freedoms of individuals in order to respect the principle of data minimisation taking into account the objective pursued.** In other words, if a technique performs the same function/allows to achieve the same result with less personal data, it must be preferred.

In particular, machine learning methods require generally the use of a very large amount of data. In order to ensure compliance with the principle of proportionality and data minimisation, the use of these technical solutions must therefore be justified. If there is a method that does not use machine learning and serves the objectives pursued, it must be preferred. Thus, the use of deep learning must be justified and should therefore not be systematic.

Example: To ensure the safety of employees, a supplier wants to delineate a hazardous area not to be crossed in his warehouse. He wants a solution that detects the presence of a person in this area and triggers an audible signal to warn the person of the danger. The use of an infrared motion detector should be preferred instead of an enhanced camera, as this solution does not collect people’s image, in accordance with the principle of data minimisation and data protection by design.

Semantic analysis of the content of a text could be carried out by a neural network based on annotated textual data, by an ensemble method such as a forest of decision-making trees or by an unsupervised algorithm, such as a clustering algorithm.

At the model training stage , account should also be taken of any uncertainty about the performance of a given architecture: compliance with the principle of minimisation shall be assessed on the basis of available scientific knowledge.

According to the advances in the concerned field, this reflection must be based on several factors for each of the architectures under consideration. This technical analysis can be done through:

- **a state of the art,** by means, for example, of:
    - a study of scientific literature (census and study of academic or private publications, specialised conferences, etc.);
    - a survey of the practices followed by professionals in the field: computer code open sourcing (including by placing it under a free license) of certain players in the sector may help to compare techniques;
    - the solicitation of the community (online competitions or challenges, online forums, conferences and dedicated meetings, etc.);

- a comparison of the results obtained after the implementation of several architectures in the form of a 'proof of concept';

- a comparison of the results obtained from the use of an existing and pre-trained model (which may possibly require fine-tuning) and a model developed by the supplier.

While the choice of AI models and algorithms may limit data collection, other design choices should be taken into account, notably with regards to the privacy by design principle. The choice of the training protocol used, in particular, may make it possible to limit access to the data only to authorised persons, or to give access only to encrypted data. Two techniques, applicable in certain situations, are particularly interesting:

- Decentralised training protocols, such as federated learning, to which a LINC article has been dedicated. This technique makes it possible to train an AI model from several datasets kept separately and, thus, for each party in the chain to keep their hands on their data. However, this technique has certain risks, concerning the security of decentralised datasets, as well as trust between actors among whom a malicious actor could conduct a poisoning attack for example.

- **The resources offered by cryptography.** Recent scientific advances in the field of cryptography can provide strong safeguards for data protection. Depending on the use cases, it may, for example, be relevant to explore the possibilities offered by secure multiparty computation, or homomorphic encryption. The techniques used in this field make it possible to train an AI model on data that remains encrypted throughout the learning process. However, they remain limited in that they cannot be applied to all types of models and because of the additional computational cost they induce. In addition, some of them, such as homomorphic encryption for training neural networks, are still under development. As technical developments are frequent in this area, it is advisable to keep an active watch on this subject.

This list of measures is not exhaustive, additional measures could be cited such as the use of a trusted execution environment, differential privacy applied during the training phase or machine learning. More generally, due to the rapid evolution of the technology, it is recommended to conduct a technological watch on the privacy practices applicable when developing AI systems.

The selection of strictly necessary data

The principle

The **principle of minimisation** provides that personal data must be **adequate, relevant and limited** to what is necessary for the purposes for which they are processed. Particular attention must be paid to the nature of the data and this principle must be applied in a particularly rigorous manner when special categories of data are processed (within the meaning of Article 9 GDPR).

In practice

The principle of minimisation does not mean that it is forbidden to train an algorithm with very large volumes of data: it involves having a reflection before training so as not to resort to personal data that would not be useful for the development of the system. In order to identify the personal data necessary, four dimensions should be taken into account:

- **Volume:** number of data subjects, historical depth, accuracy of data, distribution of data according to situations and populations, etc. It may be justified, for example, by the limited computing capacities of the servers used for training, the needs in terms of representativeness of the dataset, the practices commonly accepted by the scientific community, a comparison of the results obtained by varying the volume of data, a statistical analysis demonstrating that a minimum amount of data is necessary to achieve significant results, etc.;

- **Categories:** age, gender, face image, social network activity, etc. The presence of special categories of data or highly personal data should be examined and justified (see how-to sheet 3, Determining the legal qualification of stakeholders). This analysis may be based on the need to train the model on counterfactual data (likely to give rise to false positives in practice), a study of the usefulness of the data categories concerned (see box below), etc. Among these categories of data, preference should be given to the least intrusive format without loss of information for the objective pursued, e.g. age or age range rather than a full date of birth;

- **Typology:** real, synthesised, augmented, simulation data, anonymised or pseudonymised data, etc.;

- **Sources:** as explained in the how-to sheet 3, Determining the legal qualification of stakeholders, identification of the data sources that are envisaged to be used, whether initial collection or re-use (data available in open source, previously collected by the provider or from data providers).

Although data selection is a generally necessary phase in order to design an AI system based on quality data, in some cases and in the alternative, it may be possible to process a dataset indiscriminately. The necessity will then have to be justified.

In addition to taking into account these technical dimensions, **particular attention will have to be paid to the nature of the data within the meaning of the GDPR**, and in particular in the case of sensitive or highly personal data.

Please note:

Issues relating to **data distribution and representativeness** should also be addressed at this stage. They are essential in order to **minimise the risk of discriminatory biases.**

Linked to this question, lies the one about the inclusion of "true negative" data in the training dataset (in particular for test and validation in order to verify the absence of certain edge cases).

As these questions are particularly important, a dedicated how-to sheet will later be published.

The validity of design choices

At the end of the previous three stages, design choices are theoretically validated and data collection can begin. In order to confirm the design choices quantitatively and qualitatively, several measures are recommended as good practice.

Conducting a pilot study

The objective of the pilot study is to ensure that the choices of a technical nature and those relating to the types of data identified are relevant. To do this, small-scale experimentations can be carried out. Fictitious, synthetic, anonymised or otherwise personal data collected in accordance with the GDPR may be used.

Examples:
The use of data from social networks on the personal pages of persons who gave their consent to the collection of their data.
This type of experimentation does not always offer a representative view of the activity encountered on social networks, but it can be adapted to certain use cases such as the identification of hate content or the study of advertising targeting on these networks. This practice is beneficial because it offers a much higher level of transparency than certain practices such as web scraping.

The design of a film recommendation system
An organisation may collect from voluntary users the list of films viewed over a week and those viewed in the following days, either by declarative data or by collecting their viewing history on dedicated sites. It can conduct its pilot study on the data thus collected by anonymising the identifiers of each user.

Solicitating an ethics committee

The association of an ethics committee with the development of AI systems is a good practice to ensure that ethical issues and the protection of human rights and freedoms are taken into account upstream.

The ethics committee may have several tasks:

- the formulation of opinions on all or part of the organisation's projects, tools, products, etc. which may be subject to ethical problems;
- the facilitation of reflection and the elaboration of an internal doctrine on the ethical aspects of the development of AI systems by the organisation (e.g. what conditions for subcontracting);
- the creation of guidance regarding collective and individual attitudes and the recommendation of certain principles, behaviours or practices.

The composition and role of this committee may vary depending on the situation, but several good practices are recommended. The ethics committee should:

- **be multidisciplinary:** the profiles of the members of the committee – employees of the organisation and/or external persons – must be diversified. Members contribute to the committee's missions and can update issues that the development teams had not considered. A good practice is to assign certain committee seats to the employees of the organisation. In addition, the diversity of the members of the committee in terms of gender, age and ethnic and cultural origin is strongly encouraged;
- **be independent:** the opinions delivered by the committee may have important implications, for example for the commercial management of a company and thus favour or disadvantage some of its projects. Thus, the peoples in the committee must not be motivated by any gain (whether financial or other) to be derived from the decisions produced. Similarly, when employees sit on the committee, decisions rendered must not have consequences for them;
- **have a clearly defined role**: in order to ensure the systematic integration of the committee, a procedure must be established to determine the conditions under which the committee meets and must be associated. Depending on the situation, the committee may simply be advisory or adopt binding opinions: both approaches have advantages and disadvantages. If the committee delivers binding opinions, its inclusion in corporate governance must be particularly well defined in terms of the body's statutes, in order to avoid its instrumentalisation. If the committee is advisory, its impact must be guaranteed, in particular by ensuring mandatory referral according to precise criteria and wide transparency of its opinions, at least within the organisation and possibly other measures such as the obligation for the project owner to reply in writing to the committee's comments;
- **be notified:** the committee is encouraged to monitor technological and usage developments, document its opinions and share its knowledge. The risks associated with the use of AI evolve with technical development and new uses in this field, and it is necessary to keep a watch, in particular through the academic literature and publications of entities competent in this field (such as the Défenseur des Droits, or the French National Advisory Ethics Council for Health and Life Sciences). The dissemination of acquired knowledge will support advice and spread good practices.

In the case of the development of an AI system, the opinion of the ethics committee could be sought on several issues:

- Do the data used for development meet the ethical criteria of the organisation?
- Could the intended operational uses for the AI system have serious individual or societal consequences? Can these consequences be avoided? Can these operational uses be excluded?
- Could the potential misuse of the AI system (whether voluntary or accidental, in particular for open source models) have serious consequences for people or society? What measures would prevent them?
- Are the technical choices sufficiently controlled by the body (in the case of the use of radically new approaches)?
- Are transparency measures sufficient for the exercise of the rights of persons or to enable them to exercise a possible remedy?
- Are the risks of discriminations that may result from the use of the system identified and have the necessary means been put in place to avoid their occurrence?
- Is the organisation structured in such a way as to prevent risks by design (whether as regards non-discrimination, data protection, copyright protection, computer security, etc.)?

Depending on the size of the organisations and the way they are structured, it is not always possible to set up an ethics committee. Nevertheless, it is essential that such reflections can be carried out to support the development of AI systems. The appointment of an 'ethics adviser' may be an alternative which allows these questions to be taken into account.

< Previous : Carrying out a data protection impact assessment if necessary

Table of content

Next : Taking data protection into account in data collection and management >

Taking data protection into account in data collection and management

07 juin 2024

The development of an artificial intelligence system requires rigorous management and monitoring of training data. The CNIL details how data protection principles relate to training data management.

Once the data and its sources are identified, the AI system provider must implement the collection and create its dataset. To this end, it is necessary **to incorporate the principles of privacy by design from.**

Collection

The collection of data is accompanied by various checks and procedures depending on the modalities and sources of data. Technically, the aim is to ensure that the data collected is relevant in view of the objectives pursued, and thus to ensure compliance with the principle of minimisation.

Collection of data by web scraping

If the data controller re-uses publicly accessible data extracted from websites with web scraping tools, it must in particular ensure that the data collection is minimised, in particular by trying to:

- limit data collection to freely accessible data;
- define, prior to the implementation, precise collection criteria;
- ensure collection of relevant data only and deletion of irrelevant data immediately after its collection or as soon as it is identified as such (when this identification is not possible at the time of the collection).

Data cleaning, data identification and privacy by design

Data cleaning

Data cleaning helps in creating a quality training dataset. This is a crucial step that strengthens data integrity and relevance by reducing inconsistencies, as well as the cost of training. Specifically, it consists in:

- correcting empty values;
- detecting outliers;
- correcting errors;
- eliminating duplicates;
- deleting unnecessary fields;
- etc.

Identification of relevant data

The selection of relevant data and characteristics is a classic procedure in the field of AI. It aims to optimise the performance of the system while avoiding under- and over-fitting. In practice, it ensures that certain classes that are unnecessary for the task are not represented, that the proportions between the different classes are well balanced, etc. This procedure also aims to identify data that is not relevant for training. Data identified as irrelevant will then have to be deleted from the dataset.

In practice, this selection can be applied to three types of objects constituting the dataset:

- **The data:** these may be 'raw', unstructured data (audio extract, image, handwritten text file, etc.) or structured (measures, observations, etc. in digital format);

- **The associated metadata:** literally "data on data", metadata provide information about the collection process (what was the acquisition process? by whom was it carried out? when? etc.), the format of the data (how should they be exploited?) or its quality;

- **The annotations and characteristics extracted from the data (or features):** descriptions attributed to the data in the case of annotations, or measurable properties extracted from the data for the characteristics (information relating to the shape or texture of an image, the pitch of sounds, the timbre or tempo of an audio file, etc.).

Several approaches can contribute to the implementation of this selection. The following is illustrative:

- The use of techniques and tools to identify the relevant characteristics (feature selection), sometimes prior to training. Principal Component Analysis (PCA) can also help identifying highly correlated characteristics of a dataset and thus retaining only those that are relevant. Many libraries such as Yellowbrick, Leave One Feature Out (LOFO) and Facets today offer implementations for selecting features.

- The use **of interactive data annotation approaches** such as active learning, which allows the user to review data performing the intended task and, where appropriate, to delete those that are not relevant. The Scikit-ActiveML library is an example of this.

- The use of data/dataset pruning techniques: this technique, discussed in several publications such as Sorscher et al., 2022 or Yang et al., 2023, reduces the computation time required for training without significant impact on the performance of the model obtained, while identifying data that is not useful for training.

Finally, in certain specific cases where the storage of data may be complex or problematic (due to the sensitivity of the data, issues related to intellectual property, etc.), the principle of minimisation can be implemented by the exclusive storage of the extracted characteristics and the deletion of the source data from which they originate.

Example: For a study of the spread of hate speech in social networks, the analysis of comments associated with a post allows a classification of user reactions, but the content of the comments itself could be removed after this analysis.

Building a training dataset for AI also often **requires data annotations**. The production and use of such data must also be subject to special data protection measures. These will be detailed in a dedicated how-to sheet.

Privacy by design

Furthermore, in addition to these necessary steps, the provider of the AI system must implement **a series of measures to integrate the principles of privacy by design.**

They must take into account the state of knowledge, their impact on the effectiveness of the training, the costs of implementation and the nature, scope, context and purposes of the processing, and the risks (of which the likelihood and severity vary) of the processing for the rights and freedoms of individuals. These measures may include:

- **Generalisation measures:** those measures are intended to generalise, or dilute, the attributes of the persons concerned by changing their respective scale or order of magnitude;

- **Randomisation measures:** these measures aim to add noise to the data in order to decrease its accuracy and weaken the link between the data and the individual.

These measures must be implemented on the data and the associated metadata.

In some cases, these measures may extend to the anonymisation of the data, in particular if the purpose does not require the processing of personal data. If data selection and management qualify as processing of personal data subject to the GDPR (and thus to these how-to sheets), further processing will no longer be affected by the regulations on the protection of personal data.

Example: An organisation wishes to build a dataset of computer code relating to industrial machines (supervisory control and data acquisition, or SCADA) from repositories of several developers. After removing any mention of the developers themselves, and then verifying the absence of identifiers or personal mentions in the comments, the dataset does not contain any personal data. It is no longer subject to data protection regulations.

For more information on these measures, see Opinion 05/2014 on G29 anonymisation techniques.

In addition, some measures, such as **differential privacy** or **federated learning**, protect data when training the AI system,. Although some of these techniques are still at the research stage, tools can be used to test their effectiveness, such as PyDP or OpenDP.

For Data

The measures depend on the categories of data concerned and must be considered in terms of their influence on the technical – theoretical and operational – performance of the system. The impact of these measures is particularly beneficial due to:

- on the one hand, their ability to reduce the consequences of a possible privacy leakage (by compromising the data contained in the dataset, or by an attack on the trained model such as a membership inference attack);
- on the other hand, the possibility of using the model in the operational phase on data subject to the same protection measures, thus offering the ability to better protect them in the operational phase.

Example: By generalising patient age information as part of the development of a diagnostic AI system, in the fields [month-year] or [year] instead of [day-month-year], the provider drastically reduces the risk of privacy leakage, without prejudice to the generalisation capacity of its system.

For metadata

Metadata may contain information that is useful to an attacker seeking to re-identify the data subjects (such as a date or place of data collection). The principle of minimisation also applies to such data and should therefore be limited to what is necessary.

For example, metadata may be required by the provider to respond to a request for the exercise of rights, as it sometimes identifies data relating to an individual. In this case, special attention should be paid to their safety.

However, if the processing of metadata is not necessary and it contains personal data, its deletion may be recommended for the purpose of pseudonymisation or anonymisation of the dataset.

For example: in the event that they reuse video protection images to constitute a training dataset, a provider that generalises the location of an image from an address to an IRIS (Ilots Regroupés pour l'Information Statistique) level may no longer be able to respond to a request for access to the data.

Monitoring and updating

Although data minimisation and data protection measures have been implemented during data collection, these measures may become obsolete over time. The data collected could lose their exact, relevant, adequate and limited character, in particular because of:

- **a possible data drift** under real conditions, that is, a discrepancy between the distribution of training data and the distribution of data under conditions of use. Data drift can have multiple causes:
  - upstream process changes, such as the replacement of a sensor, the calibration of which differs slightly from that previously installed;

    - data quality problems, e.g. a broken sensor that would always indicate a zero value;
    - the natural drift of the data, such as the variation in average temperature over the seasons;
    - the appearance of a new category in a classification problem;
    - drift due to sudden changes, such as the loss of a system's ability to detect faces following the massive wearing of masks during the Covid-19 outbreak;
    - changes in the relationship between characteristics;
    - malicious poisoning as part of continuous learning, which can for example be noticed by abnormal outcomes.

Tools exist to detect the occurrence of data drift, such as Evidently, or the Scipy Library whose statistical testing functions can be used for this purpose;

- **an update of the data**, such as a correction of the place of residence in the public profile of the user of a social network following a move;
- **the evolution of techniques,** which frequently demonstrates that a change of approach (use of a different AI system requiring a different data typology, for example) can bring better performance to the system, or that similar performance can be achieved with a smaller volume of data (as shown by the technique of few-shots learning, for example).

Thus, the system provider should conduct regular analysis to monitor the dataset. This analysis will be more extensive and frequent in situations where the above-mentioned causes are most likely to take place. This analysis should be based on:

- **a regular comparison** of data or a sample of data to source data, which can be automated;
- **a regular review** of data by staff trained in data protection matters, or by an ethics committee, responsible in particular for verifying that the data is still relevant and adequate for the purpose of the processing;
- **a watch** on the scientific literature in the field and making it possible to identify the emergence of new, more data frugal, techniques.

Data storage

The principle

Personal data cannot be stored indefinitely. The GDPR requires a period of time after which data must be deleted, or in some cases archived. This retention period must be determined by the controller according to the purpose which led to the collection of such data.

In practice

The provider must set a retention period for the data used for the development of the AI system, in accordance with the principle of storage limitation (Article 5.1.d GDPR).

Defining a retention period requires, in particular, the implementation of certain procedures described in the CNIL's practical guide on retention periods. The CNIL notes that open source datasets are constantly evolving (by improving annotation, adding new data, purging poor quality data, etc.): a storage period of several years from the date of collection must be justified.

Set a retention period for the development phase

First, the provider of the AI system will have to set a data retention period for the use made during the development of the system. During this phase, the provider processes the data for:

- the creation of the dataset, which should be limited to strictly necessary, cleaned, pre-processed and ready to be used data for training;

- the training phase, from the first training of the AI model to the test phase to determine the characteristics and performance of the finished product. During this phase, the data must be kept securely and accessible to authorized persons only. Depending on the case, this phase can last from a few weeks to several months, or on the contrary iteratively in the case of continuous learning. This duration should be defined upstream and justified (taking into account the previous experiences of the controller, the knowledge of the duration of IT developments, the human and material resources available to carry them out, etc.).

Data storage needs to be planned upstream and monitored over time. The defined retention periods must also be applied to all concerned data, regardless of their storage medium. Compliance with retention periods can sometimes be facilitated by the use of management and governance tools to define a retentionstorage period for each data and calculate the length of time that has elapsed since the date of entry into the dataset before automatically deleting them. Particular attention must therefore be paid to the traceability of any data extracted from the main dataset and stored on third-party media, for example to enable the engineers to analyse a sample on a case-by-case basis. The measures recommended in the "Documentation" section for data traceability will facilitate the tracking of data and the expected date for deletion.

Please note:

In the case of public bodies or private bodies entrusted with a public service mission, data may also be subject to specific archiving in compliance with the obligations of the French *Code du Patrimoine*.

This allows the data to be permanently stored in a public archive service according to the particular interest they present. Where the public archives contain personal data, a selection shall be made to determine the data to be retained and those, which are not administratively useful or of scientific, statistical or historical interest, to be deleted.

In any event, the data retained in the context of the definitive archiving are subject to processing for archival purposes within the meaning of the GDPR and, therefore, do not fall within the scope of these how-to sheets. In addition, the data storage period must be specified in the information notice that will be brought to the attention of the data subjects.

Set a retention period for the maintenance or improvement of the product

When the data no longer have to be accessible for the day-to-day tasks related to the development of the AI system, it should in principle be deleted. However, it can be kept for product maintenance (i.e. for a later phase of performance verification or for system improvement).

Maintenance operations

The principle of data minimisation requires keeping only the data strictly necessary for maintenance operations (by selecting the relevant data, performing pseudonymisation of the data where possible, such as blurring images for example, etc.).

These operations ensure the safety of those affected by the use of the AI system in the deployment phase, such as when the system has an effect on people, when a drop in performance could have serious consequences for people, or when it concerns the safety of a product. Thus, **the storage of training data can allow audits to be carried out and facilitate the measurement of certain biases**. In these cases, and where a similar result could not be achieved through the storage of general data information (such as documentation on the model proposed in the documentation section, or information on the statistical distribution of the data), prolonged data retention may be justified. However, this storage must be limited to the necessary data and should be accompanied by enhanced security measures.

Once the data has been sorted, it can be stored on a partitioned medium, i.e. physically or logically separated from the training data. This partitioning makes it possible to strengthen the security of the data and restrict its access to authorized personal only. The duration of the maintenance phase can vary from a few months to several years when the storage of this data carries little risk to people and the appropriate measures have been taken. In the case of data from open sources, the retention period provided by the data source shall be taken into account in determining the duration of the maintenance phase. However, this period must be limited and justified by a real need.

Improving the AI system

The data of the previously created dataset may also be required to improve the product resulting from the AI system thus developed. This purpose, for which a legal basis must be identified, must be brought to the attention of the data subjects, in accordance with the principle of transparency.

Specifically, only the data needed to improve the AI system can be extracted from their partitioned storage space.

Please note:

The possibility of extending the cycle by a new development or maintenance phase will not, under any circumstances, allow an indefinite extension of the retention period. An analysis of the duration necessary for the processing operations must be carried out systematically.

Security

The principle

The controller and its processors (if any) must implement appropriate technical and organisational measures to ensure a level of security appropriate to the risks (Article 32 GDPR).

The choice of measures to be implemented must take into account the state of knowledge, the costs of implementation and the nature, scope, context and purposes of the processing and the risks, the likelihood and severity of which vary, for the rights and freedoms of data subjects.

In practice

Thus, the provider of an AI system must, in particular, put into place the appropriate measures in order to secure:

- **the techniques used for data collection** , e.g. through encryption and robust authentication methods. It is recommended to use the means provided by the diffuser to collect data, especially when it is based on APIs. The CNIL's recommendation on the use of APIs will then have to be applied;

- **the collected data**, using methods of encryption of backups, verification of their integrity, or logging of operations carried out on the dataset in accordance with the CNIL recommendation on logging measures. A frequent risk in the development of AI systems concerns duplication of data, the quality of which should be frequently analysed . Duplication of data should be limited to the extent possible and traced where unavoidable. Dedicated tools, such as NB Defense, Octopii, PiiCatcher, or techniques, such as regular expression (RegExp) search or named entity recognition for textual data, make it possible to verify the presence of personal data in certain contexts;

- **the IT system used for the development of the AI system, e.g. by means** of authentication methods and the training of staff with access to it, and the implementation of good IT hygiene practices;

- **the IT hardware,** in particular by means of methods of restricting access to premises and by analysing the guarantees provided by the data host when this is outsourced to a provider.

Security measures specific to the development and deployment phases of AI systems will be the subject of a subsequent how-to sheet. However, the recommendations and best practices traditionally implemented in IT, such as those present on the CNIL's website, as well as the GDPR guides of the development and security team of personal data, constitute a useful reference to which the provider of the AI system can refer.

Documentation

The documentation of the data used for the development of an AI system helps ensurings the traceability requirement. It must make it possible to:

- facilitate the use of the dataset;
- demonstrate that the data were lawfully collected;
- facilitate the monitoring of data over time until it is deleted or anonymised;
- reduce the risk of unanticipated use of data;
- enable the exercise of rights for data subjects;
- identify planned or possible improvements.

In order to meet these objectives, a documentation model may be adopted, in particular where the provider uses multiple data sources or creates multiple datasets. Building on the existing models (such as those proposed by Gebru et al., 2021, Arnold et al., 2019, Bender et al., 2018, the Dataset Nutrition Label, or the technical documentation provided for in Annex IV of the AI Act), the CNIL provides below a model that can be used for this purpose, in particular where the dataset is intended to be shared. This documentation should be carried out for each dataset constituted, made available, or based on existing datasets to which a substantial changes were made. More specific documentation templates for each use case, such as the CrowdWorkSheets model, which is particularly relevant for documenting the annotation phase, may complement the proposed template.

The objectives of this documentation are to help the controller clarify its practices, to inform the users of the dataset about the conditions of its constitution and the recommendations concerning its processing, and finally, to inform people about this processing. Thus, it is recommended that this documentation be provided to users of the dataset or models it has been used to design.

It should be noted that this important documentation work can naturally feed into the data protection impact assessment (see how-to sheet 5 Carrying out an impact assessment if necessary).

> Download the documentation template

< Previous : Taking into account data protection when designing the system

Table of content