BCLP. Client Intelligent

**Insights**

# AI DEVELOPERS – MAKE SURE YOU ARE COMPLIANT WITH THE GDPR!

Apr 18, 2024

## OVERVIEW

The CNIL's newly released recommendations for AI system developers set out the regulator's expectations for the entire development process of an AI system, from design to database creation and integration, ensuring development takes place in line with both GDPR and the EU's AI Act. Intended to remind AI systems developers of the centrality of EU GDPR rules to the AI development process, the recommendations also provide a step by step guide, taking a developer from conception of the system to delivery. Each step of the design process is mapped against the relevant GDPR obligations (from defining purposes of the AI system and the legal basis for processing, to minimising and securing data and conducting data protection impact assessments). Note that these recommendations do not prescribe guidance for *users* of AI systems, but are aimed at AI system "providers" as defined in the AI Act. The CNIL is already planning to draw up a number of other recommendations relating to the development and use of AI systems and will be issuing separate guidance for users.

These are the first set of guidelines made available by an EU national regulator taking into account EU's AI Act and amplify the EDPB's current AI guidance. The CNIL also emphasises that GDPR applies even to databases established outside the EU (which do not contain personal data of individuals not located in the EU), where this personal data is reused by a controller or processor established in the EU. Although these recommendations are not binding, they represent a significant step towards ensuring the ethical and legal development of AI systems and highlight the importance of clarity in purpose and responsibility in AI system development, as well as the potential extraterritorial reach of EU GDPR.

In the wake of the adoption of the EU's AI Act, the CNIL's recommendations represent timely guidance for AI system developers on how to ensure compliance with personal data regulations in tandem with the requirements of the new EU AI Act regime.

## 1ST STEP: DEFINE THE PURPOSES OF THE AI SYSTEM

Where personal data is involved, the AI system must be developed for a documented and predetermined purpose, which purpose may depend on how the AI system is to be used in the deployment phase.

The CNIL identifies three potential scenarios:

- The proposed operational use of AI is clearly identified from the start of the development phase. When AI is developed for a single operational use, the purpose in the development phase is directly linked to that pursued by the data processing in the deployment phase.
- When the system is developed for several operational uses, the development may pursue several purposes.
- In practice, the operational use in the deployment phase is often not clearly identified in the development phase. This is the case for general-purpose AI systems. The purpose will be considered as sufficiently identified, explicit and legitimate if it refers cumulatively to:
  - The type of system developed (e.g. image generative AI system); and
  - Technically feasible functionalities and capabilities. According to the CNIL, the developer must therefore draw up a list of reasonably likely capabilities at the development stage. Otherwise, the purpose will not be sufficiently precise, and the processing may not comply with GDPR.

In addition to these *minimum* compliance requirements, the CNIL suggests best practice for developers when defining the purpose of processing more precisely will be to:

- Anticipate the potential functions which could pose a risk;
- Indicate which features are excluded;
- Specify, as far as possible, the conditions of use of the AI system.

For AI systems developed for scientific purposes, the CNIL accepts that the purposes pursued by the developer may be less precisely detailed at the outset. As per the EDPB Guidelines on consent, scientific research means a research project set up in accordance with relevant sector-related methodological and ethical standards in conformity with good practice. The CNIL's Q&A on scientific research notes that scientific purposes could include algorithm improvement.

## 2ND STEP: DETERMINING THE DEVELOPER'S GDPR RESPONSIBILITIES

Identifying the role played by an AI system developer (and therefore its liability under GDPR) will always require a case-by-case analysis. The CNIL does, however, provide some food for thought:

### DATA CONTROLLER

As a starting point, the entity initiating development of the AI system (and what constitutes the training database) to the extent it determines the purposes and means of the processing, acts as the data controller. This will also be the case when the entity uses a third-party service provider, if it provides to sufficiently documented instructions to that third party. The service provider will be a processor. Nevertheless, for many AI systems, the developer may use training data collected by a third party. The CNIL distinguishes between the data disseminator - the person who makes personal data available for re-use - and the data reuser. In principle, these two parties are independent data controllers.

### JOINT CONTROLLERS

Where an AI system's training database is fed by several data controllers for a jointly defined purpose, they may be joint data controllers. The CNIL [recommends/requires] that an agreement is concluded between the joint controllers.

## 3RD STEP: DEFINE THE LEGAL BASIS APPLICABLE TO THE PROCESSING OPERATION

In practice, the developer will have to choose the legal basis best suited to the processing.

While consent is the most appropriate basis, the situation is very different in the case of databases accessible online or *open source.* In the latter case, legitimate interest could be an appropriate legal basis for processing. Given the importance of this legal basis in the context of AI development, the CNIL is preparing guidelines on the scope of this legal basis for processing.

A contract may also constitute the necessary legal basis for data processing, however recent CJEU case law emphasizes that the data processing must be objectively indispensable for the performance of the contract. This suggests that a large online platform will not be able to rely on its contracts with users as a legal basis to state in general terms and conditions of use that it intends to re-use its users' data (provided by them, observed or deduced by the operator) to develop and improve new AI products, services and functionalities useful to its users, where this is not objectively essential for use of the online platform.

## 4TH STEP: RE-USE OF DATA

A developer must conduct additional checks if it re-uses data. A distinction is drawn between the following cases.

### THE DEVELOPER REUSES DATA THAT HE HAS COLLECTED

If the data subjects have not been informed of this re-use, the developer must check the compatibility of this further processing with the initial purpose by carrying out a "*compatibility test*" (except for re-use for statistical or scientific research purposes).

The compatibility test must be based on the following elements:

- The existence of a link between the initial purpose and the purpose of the subsequent processing;
- The context in which the personal data was collected, in particular the reasonable expectations of the data subjects, looking at the relationship between the data subjects and the controller;
- The type and nature of the data, particularly in terms of its sensitivity (biometric data, geolocation data, data concerning minors, etc.);
- The possible consequences of the subsequent processing for the data subjects;
- The existence of appropriate safeguards (such as encryption or pseudonymisation).

### THE DEVELOPER REUSES A PUBLICLY ACCESSIBLE DATABASE

In this scenario, the developer must check it is lawful to re-use the database. The developer may not re-use a database that it knows or suspects does not comply with GDPR or other rules, such as those prohibiting breaches of information system security or infringements of intellectual property rights. It must therefore check that:

- The description of the database mentions its source,
- The creation and publication of the database is not the result of a crime or misdemeanor, or has been the subject of a public conviction or sanction,
- There are no obvious signs that accessing or using the database would be unlawful (by reviewing the confidentiality policy),
- The database does not contain any sensitive data.

### THE DEVELOPER REUSES A DATABASE ACQUIRED FROM A THIRD PARTY

Again, the developer will need to check it is lawful to re-use the database.

This sharing of personal data should also be governed by a contract (to protect the developer, at least contractually, against any claims that it is not in fact lawful to use the third party database). The CNIL recommends that the contract should cover:

- The source of the data, the context of the data collection, the legal basis for the processing and the data protection impact assessment (DPIA);
- Details of the information provided to individuals (in particular, the purpose and recipients); and
- Any guarantees as to the lawfulness of this data sharing by the initial holder of the data.

Please note: if the data controller wishes to base its processing on consent obtained by a third party, it must be able to provide proof that valid consent has been obtained from the data subjects. The obligation to provide proof of consent cannot be fulfilled by the mere presence of a contractual clause committing one of the parties to collect valid consent on behalf of the other party.

## 5TH STEP: MINIMISING AND SECURING PERSONAL DATA

The GDPR minimisation principle poses a significant challenge in the AI sector, as most AI systems are usually trained on a large amount of pre-existing data (although this is not always personal data). Developers should prioritize techniques that use the least amount of personal data to achieve the desired result.

To do this, it must consider:

- The purpose of the system, by reference to the type of result expected;
- The method to be used, by selecting that which is most respectful of rights and freedoms, given the objective sought. Consequently, the use of machine learning technology is only justified if there are no other, less invasive methods;
- Selection of strictly necessary data;
- The validity of design choices;
- And on this basis, the organisation of the collection.

If the web scraping is used, data collection should be limited to freely accessible data, with precise collection criteria defined beforehand. Only relevant data should be collected and any irrelevant data should be deleted immediately.

Once the data has been collected, it is possible to design the training database, by cleaning up the data before it is included, to ensure its integrity and relevance. Measures such as generalisation, randomisation, or data anonymisation should be implemented from the system's design stage to incorporate personal data protection principles (Privacy by design).

Given the dynamic nature of AI systems, these measures should be regularly monitored and updated. Regular data reviews, comparisons with source data, and monitoring of technical developments should ensure the data remains accurate, relevant, adequate, and limited.

Finally, the data used should be documented to ensure traceability and GDPR compliance. The CNIL has issued model documentation which can be accessed here.

## 6TH STEP: DEFINING A RETENTION PERIOD

The developer must set data retention periods for: (i) the development phase (based on use made during development of the system, i.e. the creation of a database and the training of the AI solution); and (ii) the product maintenance and improvement phase. During the second phase, while, as a matter of principle, data that is no longer useful for the development of AI should be deleted, it can be kept for system maintenance or improvement in a partitioned storage space, provided that a clear retention period is specified.

Exceptionally, learning data may be retained for longer periods than those specified for the phases above if it is used to carry out audits for the purpose of measuring certain biases, provided that such retention is limited to the data required.

## 7TH STEP: REQUIREMENT FOR A DATA PROTECTION IMPACT ASSESSMENT (DPIA)

GDPR requires a DPIA if the proposed processing is likely to result in a [high/significant] risk to the rights and freedoms of individuals. Those high-risk AI systems (as defined in the EU AI Act) are presumed to require completion of a DPIA (as will general-purpose AI systems). Note that any DPIA required is in addition to the obligations in the AI Act to provide compliance documentation.

Among the criteria originally identified by the European Data Protection Board (EDPB), the AI system developer must carry out a DPIA if two of the following criteria are met:

- the data to be collected is sensitive data or data of a highly personal nature such as location data or financial data;
- it involves a large-scale collection of personal data;
- data is to be collected from vulnerable persons (e.g. minors);
- the controller proposes to cross-reference or combine data sets;
- it involves the use of innovative or new technological solutions. Note: the use of AI is not automatically considered to be an innovative use, as some AI systems are based on scientific and technical innovations already in widespread use.

Given the EDPB criteria, the CNIL has published a list of personal data processing operations for which a DPIA is mandatory. Several of these may involve use of an AI system, such as those involving profiling or automated decision-making: in these cases, a DPIA is always required.

Developers must also assess the risks posed by the mere creation and use of a database, in particular data misuse, data breach or the risk of production of discriminatory results. Where these risks exist, the developer is obliged to carry out a DPIA even if it does not meet two of the EDPB's criteria. Conversely, if two criteria are met but no risk is posed by the creation of the database, then a DPIA is not required.

## RELATED PRACTICE AREAS

- Digital Transformation & Emerging Technology
- Technology Transactions

## MEET THE TEAM


**Pierre-Emmanuel Froge**

Paris

pierreemmanuel.froge@bclplaw.com
+33 (0) 1 44 17 76 21



**Anna Blest**

London

anna.blest@bclplaw.com
+44 (0) 20 3400 4475


---