Call for input on the creation of datasets for the design of artificial intelligence systems

Summary of contributions

On July 27, 2023, the CNIL published a call for contributions on cnil.fr to feed its reflection, ahead of the publication of the fact sheets on the creation of datasets for the design of artificial intelligence systems.
The contributions fed the CNIL's work with a view to the publication of the first fact sheets on the creation of datasets for the design of artificial intelligence systems, in consultation until November, 16, 2023.

CNIL.
COMMISSION NATIONALE
INFORMATIQUE & LIBERTÉS

Summary of contributions

In order to benefit from the practical and operational expertise of AI actors, the CNIL wished to collect the contributions of all the actors concerned on several structuring points of the analysis:

- **the question of the purpose (objective), in particular for general purpose AI;**
- **methods for selecting, cleaning and minimising data, available at the state of the art;**
- **approaches to take into account data protection by default and by design;**
- **the criteria to be taken into account if the legitimate interest is the legal basis for database collection processing and configuration processing (sometimes referred to as "training") of the artificial intelligence model.**

Thus, any private or public actor concerned was invited to participate in this call for contributions, in particular through concrete examples of situations encountered. The contributions were free and there was no need to answer all the questions raised in the questionnaire.

The CNIL received replies from 9 participants to the call for contributions, including:

- **5 private companies;**
- **1 research institute;**
- **1 health institution;**
- **1 trade union of employees;**
- **1 individual.**

The private companies that responded to the call for tenders specialise in various fields, such as data anonymisation, digital marketing, AI security, digital identity and defence.

Define a purpose

In general, participants confirm the difficulty of defining a purpose for the design of an AI system in some cases.

The proposed approach based on a reference to re-use patterns, tasks and capabilities of the system seems to be welcomed overall.

Furthermore, the participants suggested that:

- the purpose is defined by a reference to the intended or possible ways of re-use of personal data, and optionally by the capacity or task of the model;
- in the case of research, the field of research is specified (demographic, medical, algorithmic audit, etc.);
- broaden the definition of the main capabilities or tasks of the model, such as the generation of synthetic content (e.g. text, image, video) or object/content identification;

Choose a legal basis and process sensitive data

One participant considered that the consent of individuals was the most appropriate legal basis (if not the only possible one), especially in the context of the design of generative AI. He also pointed out that the negotiation of collective agreements might be relevant in this case.

One participant pointed out that it was difficult to realise the balance of interests needed to mobilise legitimate interest in improving algorithms used in medical devices from data collected during care.

Participants indicated that for the processing of so-called "sensitive" data, including biometric data, consent seems difficult to implement in the case of AI.

Categorise a processing as 'scientific research' within the meaning of the GDPR

Participants suggested that:

- the use in a commercial process should not be compatible with the status of scientific research;
- the source of the funding should be taken into account when assessing whether it is scientific research;
- the publication of the model, or training and evaluation techniques in a scientific journal, should be elements to consider when identifying data processing for scientific research purposes;
- the concept of methodological contribution, such as the development of a relatively generic AI method to address a family of issues, should be taken into account in order to qualify the processing as scientific research;
- open source dissemination of models should be taken into account, open source being an essential component of AI research;
- it should be possible to remain in scientific research subject to compliance with certain criteria (publication, reproducibility, etc.), for example for a model freely accessible under an open source license, if only the support services are marketed (hardware, software, model improvement, etc.);
- the classification of the system as a medical device in the field of health, or the award of an intellectual property right, should be used to distinguish the research purpose from the commercial purpose in law.

Some participants suggested that regulatory sandboxes should be put in place to bridge the gap between research and commercial model requirements to guide the transition from research to commercial use.

Minimising data

- **For data selection**

Participants suggested:

- The use of active learning after collection and before annotation;
- Identification of irrelevant data after collection, as it will have been possible to assess their usefulness for training, and any unnecessary storage or transfer can then be avoided;
- Promote techniques to reduce data accuracy, while taking into account the impact of these techniques on model training and bias reduction capabilities;
- To study the impact on performance obtained as a result of the deletion of variables or a reduction in the volume of data;
- Use techniques for selecting relevant variables and feature engineering techniques of the main component analysis;
- Measuring the impact of minimisation on the measurement and reduction of bias;
- To prioritise the synthesis of data to avoid collection (especially for ex-post studies as for comparing models or for validating them).

- **To determine the most suitable AI development process (and associated data)**

Participants noted several advantages and disadvantages to each of the organisational models for the development of an AI system, including:

- For internal development: this model allows for more control and customisation but requires more expertise and specialised resources;
- For outsourced development: this model may be more efficient and cost-effective but may lead to a loss of control;
- For the use of pre-trained models or transfer learning: this model may turn out to be faster and require less collection, but requires the basic model to match the task being sought.

In addition, some participants pointed out that the most appropriate criteria appear to be machine learning performance, speed, calculation cost (in capital investment and time), system maintainability (including debugging) and system robustness.

Building a quality database

- **For data quality**

Participants recommended:

- the use of random examinations (in particular to detect the presence of AI-generated data), cross-validation of annotations (annotation in parallel and independent by two persons, measurement of the inter-annotative agreement);
- the use of automated processes, in particular to detect data drift, or to verify the compatibility of data from several sources.

- **For the reduction of biases**

Participants recommended the use of data re-sampling, equity-sensitive algorithms, re-weighting, periodic retraining, or adversarial learning.

They also expressed the need for recommendations regarding the processing of sensitive data for the measurement of bias.

- **For representativeness**

Participants recommended:

- to study the statistical distribution of data (average, median, standard deviation);
- the use of sampling techniques (random, stratified or systematic sampling, on/subsampling), validation of generalisation (e.g. k-fold validation method);
- the use of the synthesis of data (although these cannot be used in large volumes because they are generally not sufficiently representative of the actual parameters);
- taking account of the need for representativeness in the follow-up to the exercise of rights.

Protecting data

- **For the retention of data**

Participants highlighted the need to adapt the data retention period to the need for model retraining.

- **For data protection by design and by default**

Several recommendations were put forward by the participants:

- The use of certain techniques, in particular differential confidentiality, multi-stakeholder calculation, federated learning and data clean rooms (accompanied by appropriate governance measures);
- The use of filters on inputs and outputs provided during design to verify that only relevant and appropriate data are processed;
- The use of match tables for the pseudonymisation of contact data (surname, first name, e-mail address) linked to biometric data;
- The implementation of a governance dedicated to the management of personal data and biometric data and its documentation;
- The establishment of a mechanism for tracing the data and the consent which authorised its collection, both to facilitate compliance with the rights provided for in the GDPR and intellectual property rights, including by means of a watermarking;

- Anonymisation in cases where this is possible, although this is not the case in some areas (such as genomics or responses to a chatbot);
- The implementation of anonymisation or pseudonymisation as soon as possible, and in particular by the provider carrying out the collection where practicable;
- Local data processing for model training or limited extraction of pseudonymous or anonymous information resulting from algorithmic processing, such as representations of data in latent space, in particular on websites.

One participant, however, pointed out that it could be difficult for an organisation to verify whether processes such as differential confidentiality could be characterised as anonymisation, both because of the legal risk in the event of an error of assessment and a lack of technical competence (in particular where these processes are proposed by a provider).

➢ **For safety**

Participants suggested:

- The use of certain techniques: reliable execution environments or secure enclaves (although there is a risk since the actors providing these services could cross-check the data provided to them and these devices sometimes have a high cost);
- The use of penetration tests (fictitious attacks) to validate security;
- The encryption of data at rest and in transit;
- Regularly evaluate the safety of models against attacks (exfiltration attacks, model extraction attack).