HUGGING FACE

Questionnaire on the application of the GDPR to AI models

https://www.cnil.fr/fr/webform/questionnaire-sur-lapplication-du-rgpd-aux-modeles-dia-questionnaire-application-gdpr-ai-models

With this questionnaire, the CNIL asks concerned stakeholders, whether they are providers, deployers or end-users of AI systems, whether their rights and freedoms may be affected by these questions, or whether they have a "theoretical" knowledge of the subject. The latter can be technical as on attack methods, on risk analysis, etc., or legal, on the compliance of actors, on the articulation with other regulations, etc.

*In case you have a **practical use case** to which the following questions could apply, we invite you to answer with concrete elements and to indicate as much as possible the specific means and practical conditions relating to your situation. These elements will make it possible to distinguish good practices, risks and constraints relating to each of the situations, in order to take them into account in the recommendations that will be produced later by the CNIL on this subject.*

Sur les risques de réidentification / The risks of re-identification

As set out above, the risk of re-identification of individuals whose personal data has been used to train a model exists in some cases. However, the corresponding threats are multifaceted and some could be excluded in some situations, while others might seem particularly likely. As described above, re-identification manifests itself in practice through regurgitation or extraction of personal data (including membership inference), although new risk categories may emerge.

What threats can lead to a re-identification of individuals from the trained model? For each of these threats, could you describe:

- the source of the risk (nature of the attacker when it is of malicious nature, motivations, financial and technical resources, level of access to the data, to the model, etc.)
- the objectives pursued

The risk of re-identification from a trained model is inherently linked to the exposure of sensitive or personal data contained in its training set. This includes both pre-training data obtained from publicly accessible web sources and user interaction data collected by

deployers of models [1]. This risk is more significant when the model has been trained on datasets that include personal data without proper anonymization or obfuscation techniques. If training data is not sufficiently documented or not accessible, it is more difficult to establish the risk of the re-identification of training data. The following can lead to re-identification:

(1) Membership inference attacks [2]: These attacks exploit the ability of an adversary to determine whether a particular data point was included in the model's training data. The objective is to confirm whether specific records were part of the training data, which could expose sensitive information or allow further identification of individuals.

(2) Model inversion attacks [3]: Model inversion attacks allow an adversary to reconstruct input features (e.g., a face) from model outputs, thus revealing sensitive attributes about the individuals whose data was used in the training process. The objective is to infer sensitive information about individuals whose data may have been used in training the model, potentially compromising their privacy.

(3) Data extraction via prompting: Attackers can apply so-called jailbreaking [4] techniques to extract sensitive information by prompting a model with thought-out questions. This has been shown to be an effective technique to extract training data and personal information from closed sourced models which are used for downstream tasks, such as ChatGPT [5].

[1] Mireshghallah et al.: Trust No Bot: Discovering Personal Disclosures in Human-LLM Conversations in the Wild https://arxiv.org/abs/2407.11438
[2] Shokri et al.: Membership Inference Attacks against Machine Learning Models https://arxiv.org/abs/1610.05820
[3] Fredrikson et al.: Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures https://dl.acm.org/doi/abs/10.1145/2810103.2813677
[4] Wei et al.: Jailbroken: How Does LLM Safety Training Fail? https://proceedings.neurips.cc/paper_files/paper/2023/hash/fd6613131889a4b656206c50a8bd7790-Abstract-Conference.html
[5] Li et al.: Multi-step Jailbreaking Privacy Attacks on ChatGPT https://aclanthology.org/2023.findings-emnlp.272.pdf

What factors may have an impact (positive or negative) on memorisation during model training (including factors that may increase the risk of memorisation, regurgitation or the ease of an attack)?

The size and nature of the training data [1] as well as the data distribution (e.g., outliers or rare tokens) [2] have a significant impact on memorisation of the model. Further, the model architecture and training regime can have an impact on the memorisation during model training.

[1] Carlini et al.: Extracting Training Data from Large Language Models https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting
[2] Hartley et al.: Neural networks memorise personal information from one sample https://www.nature.com/articles/s41598-023-48034-3

What factors can have an impact (positive or negative) on the likelihood of regurgitation or data extraction (in particular those relating to the motivations of the attacker, the ease of conducting the attack, the existence of alternative attacks, etc.)?

The likelihood of regurgitation or data extraction from a trained model is influenced by various factors related to both the model itself and context in which it is deployed. At the core of the issue is the presence of personal or other sensitive data in its training sets. Other relevant aspects are the attacker's resources, the technical feasibility of executing the attack, and the availability of alternative methods for data extraction.

Machine learning models can memorise sensitive data, particularly when personal data is present in the training data without sufficient anonymization techniques [1]. If the training data has been carefully curated and anonymized to remove personal data and sensitive information, the risk of regurgitation decreases significantly. While not perfect or comprehensive, data anonymization techniques such as differential privacy, k-anonymity, and tokenization can help in minimising the presence of identifiable information. If personal or sensitive data remains in the training set without adequate protection, the likelihood of the model memorising and regurgitating such information increases. In cases where sensitive information like addresses or medical records is present, the risk is particularly high, as attackers may attempt to extract this data by prompting the model in specific ways. Without access to the training data, it is hard to compare different models towards how well they deal with personal data as we can't estimate the amount of data contamination in the source data.

Deployment-level and fine-tuning interventions, including Reinforcement Learning from Human Feedback (RLHF), also have a role to play in mitigating privacy risks by lowering the likelihood that a model produces specific personal data when queried for it. These interventions have been shown to be susceptible to techniques such as jailbreaking [2] however, and should not be understood to constitute a robust technical solution by themselves.

Openly shared models also have their own risks and benefits with respect to handling risks of regurgitation or data extraction. Access to a model's weights makes studying its behaviours and encoded information easier in general, which may include surfacing encoded personal data. Identification of privacy risks is also easier for open models, especially since external stakeholders can more easily stress-test them and reproduce settings that lead to data regurgitation, facilitating the development of privacy-preserving interventions at all levels of the development chain.

[1] Song et al.: Machine Learning Models that Remember Too Much
https://arxiv.org/abs/1709.07886
[2] Li et al.: Multi-step Jailbreaking Privacy Attacks on ChatGPT
https://aclanthology.org/2023.findings-emnlp.272.pdf

What could be the consequences of regurgitation or data extraction for people whose data was used for training?

Consequences of regurgitation or data extraction for people whose data was used for training can range from privacy violations to potential defamation.

If an AI model regurgitates personal data, such as names, addresses, phone numbers, or even more sensitive data like medical or financial records, it can lead to privacy breaches. This exposure might occur unintentionally when the model is queried, or in cases of targeted attacks aiming to extract this data.

Further, AI models are known to “hallucinate” - generating inaccurate or fabricating information [1]. For example, AI models, specifically large language models (LLMs), have fabricated fals claims of sexual assault [2,3]. If the model inaccurately regurgitates accurate personal data paired with made up sensitive or defamatory information, the individuals involved can suffer significant harm.

At the ecosystem level, sensitive data leaks also erode trust in AI research and development activities. While some of the larger commercial developers of AI systems have access to sufficient quantities of proprietary data to compensate for this phenomenon, open datasets and open research efforts are more likely to suffer from this decreased trust [4].
[1] Huang et al.: A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions https://arxiv.org/abs/2311.05232
[2] https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/
[3] https://www.cityandstateny.com/politics/2024/04/meta-ai-falsely-claims-lawmakers-were-accused-sexual-harassment/396121/
[4] Longpre et al., Consent in Crisis: The Rapid Decline of the AI Data Commons https://arxiv.org/abs/2407.14933

What factors can have an impact (positive or negative) on the severity of the consequences for people?

One of the most significant factors is the nature of the information itself. If the leaked data is already public or of low sensitivity, such as a public figure’s name or widely known facts about their work, the potential impact may be limited. However, when private or sensitive data is regurgitated, such as medical records, legal information, or personal communications, the consequences can be severe. Furthermore, the situation worsens when the model generates false or misleading information as detailed above. Even for public figures, such misinformation can have a detrimental effect.

In the event that the regurgitation or extraction risk analysis of the model is only required in certain cases, a list of criteria to help identify these situations seems necessary. The following list could be used to identify cases where the risks of regurgitation and extraction are sufficiently low (these cases are characterised by the fact that none of the criteria below is

met). Do you consider the criteria in this list to be relevant? Does this list seem to you to be exhaustive?

- The identifying nature of the data, such as the presence of surnames, first names, pictures of faces, addresses or dates of birth in the training set;
- The presence of rare data or outliers;
- The duplication of training data points;
- The number of parameters of the model in relation to the volume of data;
- The functionality of the model (generative or discriminative for exemple);
- Overfitting (e.g. a metric showing better performance on training data than on validation data).

The criteria on this list are generally relevant. Currently missing are criteria, which focus on the context of the data. For example, the sensitivity of the data: Beyond personal identifiers, other sensitive information (e.g., financial data, health records) could pose regurgitation risks, even if not explicitly listed. Further, the way the model is trained, and its architecture, might influence the likelihood of memorisation or data extraction risks, such as the influence of data introduced in reinforcement learning or fine-tuning [1].

[1] Nasr et al.: Scalable Extraction of Training Data from (Production) Language Models https://arxiv.org/abs/2311.17035

Sur les techniques permettant d'analyser les risques de régurgitation et d'extraction de données personnelles / Techniques for analysing the risks of regurgitation and extraction of personal data

As mentioned above, techniques to analyse the risk of regurgitation and subsequent extraction are still not mature and do not usually provide formal guarantees. However, these practices may be sufficient in some situations to reduce the risk to an acceptable residual level. This section aims to identify existing techniques, assess their practicality, and their theoretical results.

What techniques are relevant for the analysis of the risks of regurgitation and data extraction? These measures may seek to assess the risk of memorisation by the model, regurgitation or of a specific attack (by membership inference, reconstruction, etc.).

While the outlined assessment methods are valuable for understanding privacy risks, evaluations should be broader, incorporating social impact assessments [1] that consider the societal implications of regurgitation and data extraction.

[1] Solaiman et al.: Evaluating the Social Impact of Generative AI Systems in Systems and

Society https://arxiv.org/abs/2306.05949

Can a threshold on the risk of regurgitation or data extraction as measured by the techniques listed in the previous question be set in order to consider the processing of the model as a processing of personal data? How to set this threshold? Do you consider any other criteria to be relevant?

There should be different thresholds set depending on the risks inherent in different deployment settings of a model - for example, a model that is widely deployed with easy access to an API raises different questions from a model used for internal data analysis. Factors such as the nature of the data (e.g., health records vs. publicly available information), the potential for harm if data were to be regurgitated, and the purpose for which the model was trained should all inform the threshold-setting process. How the data can be extracted, and whether it was part of the original training dataset should be used as a baseline to decide how to define thresholds.

Among these techniques, do you identify any difficulties in their implementation?

Implementing techniques to assess the risks of regurgitation and data extraction in AI models that are only accessible through APIs can be challenging due to the inherent opacity of these models and the lack of access to original training datasets. Without this access, researchers face significant hurdles in identifying which data points to target during attacks, as well as in designing effective probing methods. For example, membership inference attacks require a clear understanding of the model's training data distribution to be effective, and when that data is hidden, the likelihood of success diminishes [1]. Additionally, the evolving nature of model architectures further complicates these assessments, as changes in training techniques or data preprocessing can affect vulnerability levels, making it difficult to establish consistent methodologies across different models.

[1] Shokri et al.: Membership Inference Attacks against Machine Learning Models https://arxiv.org/abs/1610.05820

Do you consider that the techniques you have identified or referred to as examples above are sufficient to assess the risks of regurgitation and data extraction? Do they make it possible to determine when these risks are zero or sufficiently low, i.e. when a model is anonymous? Please justify your answer.

Currently, the techniques available for assessing the risks of regurgitation and data extraction in models are insufficient to guarantee that these risks are sufficiently mitigated, particularly in terms of achieving model anonymity. While methods such as membership inference and data reconstruction attacks can provide insights into potential vulnerabilities, they are not foolproof. Research has shown that even models adopting state-of-the-art privacy mitigations can inadvertently memorise and regurgitate sensitive data, highlighting the ongoing challenges in

ensuring anonymity (see previous answers).

Furthermore, as model architectures evolve and new training techniques are developed, the landscape of risks associated with regurgitation and data extraction can change significantly. This variability means that what might be considered an acceptable risk threshold today may not hold in the future, necessitating continuous research and adaptation of assessment methods.

In which situations do you consider it necessary to repeat the analysis of a model?

When models are retrained or when new data is introduced in any form (e.g., through fine-tuning or RHLF), these analyses should be repeated to ensure that the data is not contaminated by the newly introduced data.

Some analyses of the model should be repeated whenever it undergoes retraining or when new data is introduced, such as through fine-tuning or reinforcement learning from human feedback (RLHF). As new data is introduced, the model might contain new sensitive data, or the models' output in relation to previously seen data might change, potentially increasing the risk of regurgitation or data extraction.

Sur l'application du RGPD à un modèle d'IA ayant mémorisé des données personnelles / The application of the GDPR to an AI model that has memorized personal data

If it is established that a model has memorised personal data, under what conditions should the mere possession of the model be treated as a retention of personal data subject to the GDPR:

- only in the case of a generative AI model capable of regurgitating the data when used?
- also when the data is not regurgitated in the normal operation of the model but can be extracted by an attack (model inversion, inference attack, etc.)?
- according to other criteria? Please specify

Memorisation in AI models is "stochastic"; a piece of information is typically understood to be memorised if an AI system leveraging the model has a high likelihood of generating media that encodes the information. This makes it difficult to quantify when and under what conditions possession of a model should be treated as retention of personal data, especially given that the likelihood of generation also depends on the specific deployment setting,

including input and output processing software. Even in the case of attacks, regurgitation of training data may be indistinguishable from model inferences about the likely value of personal characteristics.

While model developers may be required to provide sufficient documentation about privacy risk to enable deployers to properly address them, equating possession of a model that can potentially generate personal data with retention of data subject to the GDPR without a well-defined evaluation methodology and threshold for assessing memorisation risks severely hindering the open sharing and development of models, including those with limited privacy risks.

Sur la responsabilité de l'analyse du caractère anonyme du modèle / Responsability for analyzing the anonymity of the model

If the user were to conduct the analysis of the anonymity of the model, what information and resources would they need? Do they generally have access to these information and resources in your opinion and, otherwise, who would grant this access? Do you think it is possible for the user to carry out the analysis of the anonymity of the model?

To effectively analyse the anonymity of a model, users would require access to detailed information about the training data, ideally including a comprehensive summary or even full access to the training dataset. This information should be part of the EU AI Acts requirements for a summary of the training data [1] as it is crucial for understanding the origins of the generated outputs—specifically, distinguishing between content that is synthesised based on learned patterns and content that is directly regurgitated from the training data. A proposal for the template of the training data summary includes such reporting on personal data [2].

In the case of closed source models, i.e., models only accessible through an API, access to this information is often restricted to developers and researchers within the organisation that created the model. Consequently, the feasibility of users independently conducting a thorough analysis of a model's anonymity is limited. Without sufficient transparency from model providers, the analysis becomes challenging, and any conclusions drawn may be based on incomplete or speculative information.

[1] The EU AI Act requires GPAI model providers to publish "a sufficiently detailed summary of the content used to train the general-purpose AI, in accordance with a template provided by the AI Office" (Article 53(1)d). The purpose is to "facilitate the exercise and enforcement of rights by parties with legitimate interests, including copyright holders" (Recital 107).
[2] OpenFuture: "SUFFICIENTLY DETAILED SUMMARY" — V2.0 OF THE BLUEPRINT FOR GPAI TRAINING DATA: A proposal for the AI Act's "sufficiently detailed summary" of content used to train GPAI models
https://openfuture.eu/blog/sufficiently-detailed-summary-v2-0-of-the-blueprint-for-gpai-training-data/

Do current practices seem to you to allow the user to carry out the analysis on their own if this was required of them, possibly by asking the provider to supply them with certain information? In all cases (open source, off-the-shelf purchase, subcontracting, etc.)? Do current practices seem to you to allow the provider to issue the results of the analysis to the users (in both cases of an anonymous and a non-anonymous model)?

Currently, the ability for users to conduct their own analysis of model anonymity largely depends on the availability of transparency and access to information from the model provider. In the case of open source models, where the full training dataset or a sufficiently detailed summary is made publicly available, users can assess which training data may be copied or regurgitated by the model [1,2]. This level of transparency is essential for users to conduct informed analyses, as it enables them to evaluate potential risks associated with data extraction or regurgitation

However, in cases of off-the-shelf purchases or subcontracted models, the situation is significantly different. Providers of proprietary models often do not disclose their training data or methodologies, limiting users' ability to conduct thorough analyses independently.

Overall, without a cultural shift toward greater transparency in model development and data handling practices, users are unlikely to be able to perform meaningful analyses on their own.

[1] Open Source AI Definition by the Open Source Initiative – draft v. 0.0.9
https://opensource.org/deepdive/drafts/open-source-ai-definition-draft-v-0-0-9
[2] The Columbia Convening report on Openness in AI
https://assets.mofoprod.net/network/documents/Policy_Readout_-_Columbia_Convening_on_Openness_and_AI_Final.pdf

Sur la responsabilité du traitement du modèle / Responsability for processing the model

Is the enforcement of rights a matter for the providers alone or also for the users of the model? To what extent? What techniques would allow users to respond to requests for the exercise of rights? What techniques, on the other hand, would require a disproportionate effort? What coordination between provider and user would ensure that requests are taken into account throughout the chain of responsibility?

The enforcement of rights is a shared responsibility between model developers and model users, including to ensure that GDPR is adequately followed [1]. Some interventions, such as data documentation, respect of opt-outs, and management of personal data in training datasets, should be the responsibility of the developers. Interventions by the developer have the benefit of being propagated to all uses of a model. On the other hand, users may be responsible for deploying models in ways that minimise privacy risks in their specific contexts.

These responsibilities are complementary. Ensuring clear documentation of the models and their potential risks with regards to personal data in the training data lets users and deployers select an appropriate model. An example of this is the HuggingChat interface, in which users can select which AI model they want to use to create a chat assistant [2].

[1] See paragraph 7 of the European Data Protection Board, "Report of the work undertaken by the ChatGPT Taskforce":
https://www.edpb.europa.eu/our-work-tools/our-documents/other/report-work-undertaken-chatgpt-taskforce_en
https://www.edpb.europa.eu/system/files/2024-05/edpb_20240523_report_chatgpt_taskforce_en.pdf
[2] Interface to create a new chat bot ("assistant") on HuggingChat, where the user can select between different models (as "Model"): https://huggingface.co/chat/settings/assistants/new

En conclusion / In conclusion

Given the risks of regurgitation and data extraction from the trained model for data subjects, and taking into account existing techniques, do you consider it proportionate to require an analysis of the anonymity of the models you process? If so, which analytical techniques do you consider sufficient? If not, what constraints make this analysis disproportionate? Please justify by providing context on your own situation.

Given the risks of regurgitation and data extraction from trained models, focusing requirements on evaluation of and transparency regarding the anonymity of these models can be seen as a proportionate measure, especially when sensitive personal data is involved. Analytical techniques may include personal data detection tools, such as those Hugging Face contributed to or used on the plattform, which leverage models to identify sensitive information within datasets [1,2]. The use of such tools can help assess the level of risk associated with a model's outputs by examining the training data for identifiable information and enable stakeholders to gauge whether the model can be considered anonymous or if additional measures are needed.

[1] https://huggingface.co/blog/presidio-pii-detection
[2] personal data (also personally identifiable information (PII)) detection report widget on the right of datasets such as: https://huggingface.co/datasets/ai4privacy/pii-masking-300k

Submitted by:
Lucie-Aimée Kaffee, EU Policy Lead & Applied Researcher, Hugging Face
Yacine Jernite, ML & Society Lead, Hugging Face