# Collecter et qualifier les données d'entraînement

05 avril 2022

Respecter le RGPD lors de la collecte et constituer une base de données de qualité.

# Traiter les données d’apprentissage de manière licite en respectant la réglementation

Lors de la constitution d'une base de données d'entraînement contenant des données personnelles (que celle-ci soit créée par le fournisseur du système d'IA ou fournie par un tiers), certaines précautions doivent être prises pour assurer le respect de la réglementation. Les vérifications devront porter, en particulier, sur l’origine des données, sur la possibilité même de les réutiliser à des fins d’entrainement ou sur les mesures prises pour limiter les risques de mésusage.

Les questions suivantes pourront aider le responsable de traitement à apprécier s’il réunit les conditions pour constituer une base de données d’entraînement respectueuse de la réglementation.

☐ Les données d'entraînement font-elles l'objet d'une réutilisation (réutilisation d’une base interne ou publiquement accessible, acquisition, etc.) ou d'une collecte spécifique ?

☐ Dans le cas d’une réutilisation, la base de données a-t-elle été constituée conformément à la réglementation en matière de protection des données ?

☐ Si une base de données publiquement accessible a été utilisée, a-t-elle fait l'objet d'études, en particulier en ce qui concerne la présence de biais ?

Sur quelle base légale s'appuie le traitement des données d'entraînement ?

Si le traitement porte sur des données sensibles (santé, infraction, etc.), le traitement n’est possible qu’en mobilisant une des exceptions prévues à l’article 9 du RGPD. Sur laquelle de ces exceptions repose le traitement ?

Comment est suivie la conformité du traitement des données d'entraînement (réalisation d’une AIPD, d’une analyse des risques de réidentification, etc.) ?

☐ Les jeux de données utilisés pour l'entraînement ont-ils été constitués de façon à satisfaire le principe de minimisation ?

☐ Les données sont-elles anonymisées ?

Par quel moyen ?

☐ Sont-elles pseudonymisées ?

Par quel moyen ?

☐ Les risques de réidentification ont-ils été évalués ?

☐ Le volume des données recueillies est-il justifié au vu de la difficulté de la tâche d'apprentissage ?

☐ Les variables envisagées pour l'entraînement du modèle sont-elles toutes nécessaires ?

☐ La collecte de certaines valeurs qui ne s'avèreraient pas utiles à l'apprentissage, en particulier si elles constituent des données sensibles, pourrait-elle évitée ?

☐ Si leur collecte ne peut être évitée, pourraient-elles être supprimées ou masquées ?

# Des données brutes à un ensemble de données d'entraînement de qualité

La qualité des sorties de l'algorithme est intimement liée à la qualité de l'ensemble des données d'entraînement, quelles que soient les catégories de données concernées.

Certains critères doivent être vérifiés afin de limiter les risques d'erreur lors de l'utilisation de l'algorithme, en particulier lorsque celui-ci engendre des conséquences pour les personnes.

☐ La véracité des données a-t-elle été vérifiée ?

☐ Si une méthode d'annotation a été utilisée, est-elle vérifiée ?

☐ Si l'annotation est réalisée par des personnes, ont-elles été formées ?

☐ La qualité de leur travail est-elle contrôlée ?

☐ Les données utilisées sont-elles représentatives des données observées en environnement réel ?

Quelle méthodologie a été mise en œuvre pour garantir cette représentativité ?

☐ Une étude formalisée de cette représentativité a-t-elle été réalisée ?

☐ Si le traitement repose sur une solution d'apprentissage fédéré (*federated learning*), le caractère indépendant et identiquement distribué des données utilisées au sein des centres (condition garantissant que les informations tirées des données reflèteront les mêmes tendances sans spécificité propre à chaque centre) a-t-il été évalué ?

S'il n'est pas vérifié, quelles mesures sont prises pour y remédier ?

Dans le cas d'un système d'IA utilisant de l'apprentissage continu, quel mécanisme est mis en œuvre afin de garantir la qualité des données utilisées de façon continue ?

☐ Des mécanismes réguliers d'évaluation des risques de perte en qualité ou de changement dans la distribution des données sont-ils mis en œuvre ?

# Identifier les risques de biais et les corriger efficacement

Les risques de discriminations liées à l'utilisation d'un algorithme entraîné sur des données biaisées sont largement connus aujourd'hui. En revanche, les facteurs contribuant à ces risques restent mal identifiés et les méthodes pour les corriger sont encore expérimentales.

Il est donc nécessaire **d'inspecter rigoureusement l'ensemble des données d'entraînement** afin d'y déceler les indices de potentiels biais.

- [ ] La méthode utilisée pour la collecte des données d'entraînement est-elle suffisamment connue ?
- [ ] Des biais peuvent-ils exister du fait de la méthode utilisée ou des conditions particulières de la collecte ?

- [ ] Les données d'entraînement comportent-t-elles des données liées aux caractéristiques particulières des personnes telles que leur sexe, leur âge, leurs caractéristiques physiques, des données sensibles, etc. ?

Lesquelles ?

- [ ] Les hypothèses effectuées sur les données d'entraînement ont-elles été discutées, clairement documentées et confrontées à la réalité ?
- [ ] Une étude des corrélations entre ces caractéristiques particulières et le reste des données d'entraînement a-t-elle été effectuée afin d'identifier de possibles proxys ?
- [ ] Une étude des biais a-t-elle été effectuée ?

Selon quelle méthode ?

Si un biais a été identifié, quelles mesures ont été prises pour le réduire ?

Imprimer cette check-list

---

< Fiche précédente : se poser les bonnes questions avant de mettre en place un système d'IA

Sommaire

Fiche suivante : développer et entraîner un algorithme >

## Vous souhaitez contribuer ?

## Écrivez à ia[@]cnil.fr

Haut de page