IA : Tenir compte de la protection des données dans la collecte et la gestion des données

08 avril 2024

Le développement d'un système d'intelligence artificielle nécessite d'assurer une gestion et un suivi rigoureux des données d'apprentissage. La CNIL détaille comment les principes relatifs à la protection des données s'articulent avec la gestion des données d'apprentissage.

Une fois les données et leurs sources identifiées, le fournisseur du système d'IA doit mettre en œuvre la collecte et constituer sa base de données. Pour cela il est nécessaire d'intégrer **dès leur conception les principes de protection des données personnelles (« *privacy by design* »).**

Collecte

La collecte des données s'accompagne de différentes vérifications et démarches en fonction des modalités et sources de données. Techniquement, il s'agit de s'assurer que les données collectées sont pertinentes compte tenu des objectifs poursuivis, et ainsi d'assurer le respect du principe de minimisation.

Collecte de données par moissonnage (« *web scraping* »)

Quand le responsable du traitement réutilise des données publiquement accessibles qu'il a lui-même extraites de sites web au moyen d'outils de moissonnage (« *web scraping* »), il doit particulièrement s'assurer de minimiser la collecte de données, en tâchant notamment de :

- limiter la collecte aux données librement accessibles ;
- définir, en amont de la mise en œuvre du traitement, des critères précis de collecte ;
- s'assurer de ne collecter que des données pertinentes et de la suppression des données immédiatement après leur collecte ou dès qu'elles sont identifiées comme telles (quand le tri exhaustif n'est pas possible lors de la collecte).

Pour plus d'informations au sujet de la collecte des données publiquement accessibles : voir le projet de guide sur l'ouverture, le partage et la réutilisation des données.

Nettoyage, identification et protection de la vie privée dès la conception

Nettoyage

Le nettoyage des données permet de constituer une base d'entraînement de qualité. C'est une étape cruciale qui renforce l'intégrité et la pertinence des données en réduisant les incohérences, ainsi que le coût de l'entraînement. Concrètement, il s'agit ainsi de :

- corriger les valeurs vides ;
- détecter les valeurs aberrantes ;
- corriger les erreurs ;
- éliminer les doublons ;
- supprimer les champs inutiles ;
- etc.

Identification des données pertinentes

La sélection des données et des caractéristiques pertinentes est une procédure classique en IA. Elle vise à optimiser les performances du système tout en évitant les sous- et sur-apprentissage. En pratique, elle permet ainsi de s'assurer que certaines classes inutiles pour la tâche visée ne sont pas représentées, que les proportions entre les différentes classes d'intérêt sont bien équilibrées, etc. Cette procédure vise également à identifier les données non pertinentes pour l'apprentissage. Les données identifiées comme non pertinentes devront alors être supprimées de la base.

En pratique, cette sélection peut trouver à s'appliquer sur trois types d'objets constituant la base de données :

- **Les données** : il peut s'agir de données « brutes », non-structurées, (extrait audio, image, fichier texte manuscrit, etc.) ou structurées (mesures, observations, etc. au format numérique) ;
- **Les métadonnées associées** : littéralement « données sur les données », les métadonnées, fournissent des informations de description (quel a été le processus d'acquisition ? par qui a-t-il été réalisé ? à quelle date ? etc.), de structure (comment faut-il les exploiter ?) ou encore de qualité ;
- **Les annotations et les caractéristiques extraites de données (« *features* »)** : descriptions attribuées aux données dans le cas des annotations, ou propriétés mesurables extraites à partir des données pour les caractéristiques (informations relatives à la forme ou la texture d'une image, à la hauteur des sons, au timbre ou au tempo d'un fichier audio, etc.).

Plusieurs approches peuvent concourir à mettre en œuvre cette sélection. Citons à titre illustratif :

- L'utilisation de techniques et outils permettant **d'identifier les caractéristiques pertinentes** (sélection de caractéristiques ou « *feature selection* »), parfois en amont de l'entraînement. Des analyses de type Analyse en composantes principales (PCA - « *Principal Component Analysis* »), peuvent également aider à identifier les caractéristiques fortement corrélées d'un jeu de données et ainsi ne conserver que celles qui sont pertinentes. De nombreuses bibliothèques telles que Yellowbrick, *Leave One Feature Out* (LOFO) ou encore Facets proposent aujourd'hui des implémentations pour la sélection de caractéristiques.
- L'utilisation d'**approches d'annotations interactives de données** comme l'apprentissage actif (« *active learning* »), qui permettent une revue des données par l'utilisateur sur la base de la tâche à

accomplir et, le cas échéant, la suppression de celles qui sont non-pertinentes. La bibliothèque Scikit-ActiveML en est un exemple.
- L'utilisation de techniques d'**ablation des données d'entraînement** (« *data/dataset pruning* ») : cette technique, abordée dans plusieurs publications comme Sorscher et al., 2022 ou Yang et al., 2023, permet de réduire le temps de calcul nécessaire à l'entraînement sans impact significatif sur les performances du modèle obtenu, tout en identifiant les données peu utiles à l'entraînement.

Enfin, dans certains cas spécifiques et pour lesquels la conservation des données pourra s'avérer complexe ou problématique (en raison de la sensibilité des données, de questions liées à la propriété intellectuelle, etc.), le principe de minimisation peut être mis en œuvre par la conservation exclusive des caractéristiques extraites et par la suppression des données sources dont elles sont issues.

Exemple : Pour une étude de la propagation de propos haineux dans les réseaux sociaux, l'analyse des commentaires associés à une publication permet de réaliser une classification des réactions des utilisateurs, mais le contenu des commentaires lui-même pourrait être supprimé après cette analyse.

La constitution d'une base de données d'apprentissage pour l'IA nécessite également souvent l'annotations des données. La production et l'utilisation de celles-ci doivent également faire l'objet de mesures particulières au titre de la protection des données. Celles-ci seront détaillées dans une fiche pratique dédiée.

Protection des données dès la conception (« *privacy by design* »)

Par ailleurs, outre ces étapes nécessaires, le fournisseur du système d'IA doit mettre en œuvre **une série de mesures pour intégrer dès leur conception les principes de protection des données personnelles** (« *privacy by design* »).

Elles doivent tenir compte de l'état des connaissances, de leur incidence sur l'efficacité de l'entraînement, des coûts de mise en œuvre et de la nature, de la portée, du contexte et des finalités du traitement ainsi que des risques (dont la vraisemblance et la gravité varient) que présente le traitement pour les droits et libertés des personnes. Ces mesures peuvent inclure :

- **Des mesures de généralisation** : ces mesures visent à généraliser, ou diluer, les attributs des personnes concernées en modifiant leur échelle ou leur ordre de grandeur respectif ;
- **Des mesures de randomisation** : ces mesures visent à ajouter du bruit aux données afin d'en diminuer la précision et d'affaiblir le lien entre les données et l'individu.

Ces mesures sont à mettre en œuvre sur les données ainsi que les métadonnées qui y sont associées.

Dans certains cas, ces mesures peuvent aller jusqu'à l'anonymisation des données, et notamment si l'objectif n'impose pas de traiter des données personnelles : si les traitements de sélection et gestion des données sont des traitements de données personnelles soumis au RGPD et donc aux présentes fiches, les traitements ultérieurs ne seront plus concernés par la réglementation sur la protection des données personnelles.

Exemple : Un organisme souhaite constituer une base de données de code informatique relatif à des machines industrielles (SCADA) issues de dépôts de plusieurs développeurs. Après avoir retiré toute mention relative aux développeurs eux-mêmes, puis avoir vérifié l'absence d'identifiants ou de mentions personnelles dans les commentaires, la base de données ne contient aucune donnée personnelle. Elle n'est plus soumise à la réglementation sur la protection des données.

Pour plus d'informations sur ces mesures, voir l'avis 05/2014 sur les techniques d'anonymisation du G29.

De plus, certaines mesures permettent de protéger les données lors de l'apprentissage du système d'IA, comme la **confidentialité différentielle** appliquée durant l'apprentissage du modèle ou l'**apprentissage fédéré**. Bien que certaines de ces techniques soient encore au stade de la recherche, des outils permettent de les mettre en œuvre afin de tester leur efficacité, comme PyDP ou encore OpenDP.

Les mesures portant sur les données

Les mesures applicables dépendent des catégories de données concernées et doivent être considérées au regard de leur influence sur les performances techniques – théoriques et opérationnelles – du système. L'impact de ces mesures est particulièrement bénéfique en raison :

- d'une part, de leur capacité à réduire les conséquences d'une éventuelle perte de confidentialité des données (par compromission des données contenues dans la base, ou par une attaque portant sur le modèle entraîné tel qu'une attaque par inférence d'attribut) ;
- d'autre part, de la possibilité éventuelle d'utiliser le modèle entraîné en phase opérationnelle sur des données ayant fait l'objet de mesures de protection identiques, offrant ainsi la capacité de mieux les protéger en phase opérationnelle.

Exemple : En généralisant les informations relatives à l'âge de patients dans le cadre du développement d'un système d'IA d'aide au diagnostic, aux champs [mois-année] ou [année] à la place de [jour-mois-année], le fournisseur réduit drastiquement les risque de perte de confidentialité, cela sans préjudice sur la capacité de généralisation de son système.

Les mesures portant sur les métadonnées

Les métadonnées peuvent contenir des informations utiles à un attaquant qui cherche à réidentifier les personnes concernées (comme une date ou un lieu de collecte des données). Le principe de minimisation s'applique également à ces données, et elles devraient ainsi être limitées à ce qui est nécessaire.

Les métadonnées peuvent par exemple être nécessaires au fournisseur pour donner suite à une demande d'exercice des droits, puisqu'elles permettent parfois d'identifier les données se rapportant à une personne. Dans ce cas, une attention particulière devrait être portée à leur sécurité.

Toutefois, si le traitement des métadonnées n'est pas nécessaire et que celles-ci contiennent des données à caractère personnel, leur suppression peut être recommandée dans un objectif de pseudonymisation ou d'anonymisation du jeu de données.

Par exemple : dans le cas où il réutiliserait des images de vidéoprotection pour constituer un jeu de données d'apprentissage, un fournisseur qui généraliserait le lieu

de collecte d'une image d'une adresse à une maille IRIS pourrait ne plus être en mesure de donner suite à une demande d'accès aux données.

Suivi et mise à jour

Bien que des mesures de minimisation et de protection des données aient été mises en œuvre lors de la collecte des données, ces mesures pourraient devenir obsolètes au cours du temps. En effet, les données collectées pourraient perdre leurs caractères exact, pertinent, adéquat et limité, en raison notamment :

- **d'une possible dérive des données** en conditions réelles, c'est-à-dire d'un écart entre la distribution des données d'entraînement et la distribution des données en condition d'utilisation. La dérive des données peut avoir de multiples causes :
    - des modifications de processus en amont, comme le remplacement d'un capteur, dont l'étalonnage diffère légèrement de celui qui était précédemment installé ;
    - des problèmes de qualité des données, par exemple un capteur cassé qui indiquerait toujours une valeur nulle ;
    - l'apparition d'une nouvelle catégorie dans un problème de classification ;
    - la dérive naturelle des données, comme la variation de la température moyenne au fil des saisons ;
    - la dérive due à de soudains changements, comme la perte de la capacité d'un système à détecter des visages suite au port massif de masques lors de l'épidémie de Covid-19 ;
    - la modification de la relation entre caractéristiques ;
    - un empoisonnement malveillant dans le cadre d'un apprentissage continu, par exemple constaté par des résultats non souhaités.

Des outils existent pour détecter l'apparition d'une dérive des données, tel que Evidently, ou bien la bibliothèque Scipy dont les fonctions de tests statistiques peuvent être utilisés dans cet objectif ;

- **d'une mise à jour des données**, tel qu'une correction du lieu d'habitation dans le profil public de l'utilisateur d'un réseau social à la suite d'un déménagement ;
- **de l'évolution des techniques**, qui démontre fréquemment qu'un changement d'approche (utilisation d'un système d'IA différent nécessitant une typologie de données différente, par exemple) peut apporter de meilleures performances au système, ou encore que des performances similaires peuvent être obtenues avec un volume de données moins important (comme l'a montré la technique du « *few-shot learning* », par exemple).

Ainsi, le fournisseur du système devrait conduire une analyse régulière pour assurer le suivi de la base de données constituée. Cette analyse sera plus poussée et plus fréquente dans les situations où les causes évoquées ci-dessus sont les plus à même d'avoir lieu. Cette analyse devrait reposer sur :

- **une comparaison régulière** des données ou d'un échantillon de données aux données sources, celle-ci pouvant être automatisée ;
- **une revue régulière** des données par des agents formés aux questions relatives à la protection des données, ou par un comité d'éthique, en charge de vérifier notamment que les données sont toujours pertinentes et adéquates pour la finalité du traitement ;
- **une veille** portant sur la littérature scientifique dans le domaine et permettant d'identifier l'apparition de nouvelles techniques plus frugales en données.

Conservation des données

Le principe

Les données personnelles ne peuvent être conservées indéfiniment. Le RGPD impose de définir une durée au bout de laquelle les données doivent être supprimées, ou dans certains cas archivées. Cette durée de conservation doit être déterminée par le responsable de traitement en fonction de l'objectif ayant conduit à la collecte de ces données.

En savoir plus : Les durées de conservation des données

En pratique

Le fournisseur doit fixer une durée de conservation des données utilisées pour le développement du système d'IA, conformément au principe de limitation de la conservation des données (article 5.1.d du RGPD).

La fixation d'une durée de conservation impose notamment la mise en œuvre de certaines procédures décrites dans le guide pratique de la CNIL sur les durées de conservation. La CNIL constate que les bases de données publiées en source ouverte évoluent constamment (par amélioration de l'annotation, ajout de nouvelles données, purge des données de mauvaise qualité, etc.) : une durée de conservation de plusieurs années à partir de la date de la collecte devra être justifiée.

Fixer une durée de conservation pour la phase de développement

Tout d'abord, le fournisseur du système d'IA devra fixer une durée de conservation des données pour l'usage fait pour le développement du système. Durant cette phase, le fournisseur utilise les données pour :

- **la constitution de la base de données** limitée à celles strictement nécessaires, nettoyées, prétraitées et prêtes à être utilisées pour l'apprentissage ;
- **l'apprentissage de sa solution**, depuis le premier entraînement du modèle d'IA jusqu'à la phase de test permettant de déterminer les caractéristiques et performances du produit fini. Lors de cette phase, les données doivent être conservées de manière sécurisée et être accessibles aux personnes habilitées. Selon les cas, cette phase peut durer de quelques semaines à plusieurs mois, ou au contraire se faire de manière itérative dans le cas de l'apprentissage en continu. Cette durée devrait être définie en amont et justifiée (en tenant compte de l'expérience passée du responsable du traitement, de ses connaissances sur la durée des développements informatiques, des ressources humaines et matérielles qu'il peut mettre à disposition pour les réaliser, etc.).

La conservation des données doit faire l'objet d'une planification en amont et d'un suivi dans le temps. Les durées de conservation définies doivent par ailleurs être appliquées aux données concernées, quel que soit leur support. Le respect des durées de conservation peut parfois être facilité par l'utilisation d'outils de gestion et de gouvernance permettant de définir une durée de conservation de chaque donnée et calculant la durée écoulée depuis la date d'entrée dans la base avant de les supprimer automatiquement. Une attention particulière doit ainsi être portée à la traçabilité des données éventuellement extraites de la base principale et sauvegardées sur des supports tiers, par exemple pour permettre l'analyse d'un échantillon au cas par cas par les ingénieurs. Les mesures recommandées dans la section « Documentation » concernant la traçabilité des données pourront faciliter le suivi des données et de la date prévue pour leur suppression.

À noter :

Concernant les organismes publics ou les organismes de droit privé chargés d'une mission de service public, les données peuvent également devoir faire l'objet d'un archivage spécifique dans le respect des obligations du code du patrimoine.

Les données peuvent ainsi être versées en archivage définitif dans un service public d'archives selon l'intérêt particulier qu'elles présentent. Lorsque les archives publiques comportent des données personnelles, une sélection est réalisée pour déterminer les données destinées à être conservées et celles, dépourvues d'utilité administrative ou d'intérêt scientifique, statistique ou historique, destinées à être éliminées.

En tout état de cause, les données conservées dans le cadre de l'archivage définitif relèvent d'un traitement à finalité archivistique au sens du RGPD et, dès lors, n'entrent pas dans le cadre des présentes fiches. Par ailleurs, la durée de conservation des données doit être précisée dans les mentions d'information qui seront portées à la connaissance des personnes concernées.

Fixer une durée pour la maintenance ou l'amélioration du produit

Lorsque les données n'ont plus à être accessibles pour les tâches quotidiennes des personnes en charge du développement du système d'IA, elles doivent en principe être supprimées. Elles peuvent toutefois être conservées pour la maintenance du produit (c'est-à-dire pour une phase ultérieure de vérification des performances) ou encore à des fins d'amélioration du système).

Les opérations de maintenance

Le principe de minimisation des données impose de ne conserver que les données strictement nécessaires aux opérations de maintenance (en sélectionnant les données pertinentes, en réalisant une pseudonymisation des données lorsque c'est possible, comme en floutant des images par exemple, etc.).

Ces opérations permettent de garantir la sécurité des personnes concernées par l'utilisation du modèle en phase de déploiement, comme lorsque le système produit un effet sur les personnes, lorsqu'une baisse de performance pourrait entraîner des conséquences graves pour les personnes, ou encore lorsqu'il concerne la sécurité d'un produit. Alors, **la conservation des données d'apprentissage peut permettre d'effectuer des audits, et faciliter la mesure de certains biais**. Dans ces cas, et lorsqu'un résultat similaire ne pourrait être atteint par la conservation d'informations générales sur les données (telles que la documentation réalisée sur le modèle proposé dans la section Documentation, ou encore des informations sur la distribution statistique des données), une conservation des données prolongée peut être justifiée. Cette conservation doit toutefois être limitée aux données nécessaires, et s'accompagner de mesures de sécurité renforcées.

Une fois les données triées, elles peuvent être stockées sur un support cloisonné, c'est-à-dire séparé physiquement ou logiquement des données constitutives de la base. Ce cloisonnement permet de renforcer la sécurité des données et de restreindre leur accès aux seules personnes habilitées. La durée de la phase de maintenance peut varier de quelques mois à plusieurs années lorsque la conservation de ces données emporte peu de risques pour les personnes et que les mesures adaptées ont été prises. Dans le cas de données provenant de sources ouvertes, la durée de conservation prévue par la source des données doit être prise en compte dans la détermination de la durée de la phase de maintenance. Cette durée doit toutefois être limitée, et justifiée par un besoin réel.

L'amélioration du système d'IA

Les données constitutives de la base constituée précédemment peuvent être également nécessaires pour améliorer le produit issu du système d'IA ainsi développé. Cette finalité, pour laquelle une base légale devra être identifiée, devra être portée à la connaissance des personnes concernées, conformément au principe de transparence.

Concrètement, seules les données nécessaires à l'amélioration du système d'IA peuvent être extraites de leur espace de stockage cloisonné.

À noter :

La possibilité de prolonger le cycle par une nouvelle phase de développement ou de maintenance ne pourra, en aucun cas, permettre une prolongation indéfinie de la durée de conservation, une analyse de la durée nécessaire aux opérations de traitement devra être conduite systématiquement.

Sécurité

Le principe

Le responsable du traitement et ses sous-traitants (s'il en a) doivent mettent en œuvre les mesures techniques et organisationnelles appropriées afin de garantir un niveau de sécurité adapté au risque (article 32 du RGPD).

Le choix des mesures à mettre en œuvre doit tenir compte de l'état des connaissances, des coûts de mise en œuvre et de la nature, de la portée, du contexte et des finalités du traitement ainsi que des risques, dont les degrés de vraisemblance et de gravité varient, pour les droits et libertés des personnes concernées.

En pratique

Ainsi, le fournisseur d'un système d'IA doit en particulier prévoir les mesures adaptées afin de sécuriser :

- **les techniques de collecte des données employées**, au moyen par exemple de méthodes de chiffrement des flux et de méthodes d'authentification robustes permettant de restreindre l'accès au système d'information. Il est recommandé d'utiliser les moyens prévus par le diffuseur pour collecter les données, notamment lorsque ceux-ci reposent sur des API. La recommandation de la CNIL sur l'utilisation d'API devra alors être appliquée ;
- **les données collectées**, au moyen de méthodes de chiffrement des sauvegardes, de vérification de leur intégrité, ou encore de journalisation des opérations réalisées sur la base de données conformes à la recommandation de la CNIL relative aux mesures de journalisation. Un risque fréquent dans le développement de système d'IA concerne la duplication des données, celles-ci ayant fréquemment à être analysées pour vérifier leur qualité. Les duplications de données devraient être limitées dans la mesure du possible et tracées lorsqu'elles sont inévitables. Des outils dédiés, comme NB Defense, Octopii, ou encore PiiCatcher, ou des techniques telles que la recherche par expressions régulières, ou la reconnaissance d'entités nommées pour les données textuelles, permettent de vérifier la présence de données personnelles dans certains contextes ;
- **le système d'information utilisé pour le développement du système d'IA**, au moyen, par exemple, de méthodes d'authentification et de la formation des agents ayant à y accéder, et de la mise en œuvre des bonnes pratiques d'hygiène informatique ;
- **le matériel informatique**, notamment au moyen de méthodes de restriction d'accès aux locaux et par l'analyse des garanties apportées par l'hébergeur de données lorsque cela est sous-traité à un

prestataire.

Les mesures de sécurité spécifiques aux phases de développement et déploiement de systèmes d'IA seront l'objet d'une Fiche ultérieure. Toutefois, les recommandations et bonnes pratiques classiquement mises en œuvre en informatique, telles que celles présentes sur le site de la CNIL, ainsi que les guides RGPD de l'équipe de développement et de la sécurité des données personnelles, constituent un socle de référence utile auquel le fournisseur du système d'IA pourra se référer.

Documentation

La documentation des données utilisées pour le développement d'un système d'IA permet de garantir la traçabilité des jeux de données utilisés dont la grande taille rend généralement cette tâche difficile. Elle doit permettre de :

- faciliter l'utilisation de la base de données ;
- démontrer que les données ont été collectées de manière licite ;
- faciliter le suivi des données dans le temps jusqu'à leur suppression ou leur anonymisation ;
- réduire les risques d'une utilisation imprévue des données ;
- permettre l'exercice des droits pour les personnes concernées ;
- identifier les améliorations prévues ou envisageables.

Afin de répondre à ces objectifs, un modèle de documentation pourra être adopté, notamment dans le cas où le fournisseur a recours à de multiples sources de données ou constitue plusieurs bases de données. En s'appuyant sur les modèles existants (tel que ceux proposés par Gebru et al., 2021, Arnold et al., 2019, Bender et al., 2018, le Dataset Nutrition Label, ou encore la documentation technique prévue en annexe IV du projet de règlement européen sur l'intelligence artificielle), **la CNIL fournit ci-après un modèle qui pourra être utilisé à cet effet**, notamment dans le cas où la base de données constituée a vocation à être diffusée. Cette documentation devrait être réalisée par jeu de données lorsque ceux-ci sont constitués, mis à disposition, ou qu'ils proviennent d'un jeu de données existant auquel une modification substantielle a été apportée. Des modèles de documentation plus spécifiques à chacun des cas d'usage, comme le modèle CrowdWorkSheets, particulièrement pertinent pour documenter la phase d'annotation, pourront compléter le modèle proposé.

Les objectifs de cette documentation sont de nourrir la réflexion interne du responsable de traitement sur ses pratiques, d'informer les utilisateurs du jeu de données sur les conditions de sa constitution et des recommandations concernant son traitement, et enfin, d'informer les personnes dans un but de transparence. Ainsi, il est recommandé de fournir cette documentation aux utilisateurs du jeu de données ou des modèles qu'il a servi à concevoir.

Il est à noter que cet important travail de documentation peut alimenter naturellement l'analyse d'impact sur la protection des données.

> Télécharger le modèle de documentation

< Fiche précédente : Tenir compte de la protection des données dans la conception du système

Sommaire