Appel à contribution sur la constitution de bases de données pour la conception de systèmes d'intelligence artificielle

Synthèse des contributions

Le 27 juillet 2023, la CNIL a publié un appel à contribution sur le site cnil.fr visant à alimenter sa réflexion, en amont de la publication des fiches portant sur la constitution de bases de données pour la conception de systèmes d'intelligence artificielle.
Les contributions ont nourri les travaux de la CNIL en vue de la publication des premières fiches sur constitution de bases de données pour la conception de systèmes d'intelligence artificielle, en consultation jusqu'au 16 novembre 2023.

CNIL.
COMMISSION NATIONALE
INFORMATIQUE & LIBERTÉS

Synthèse des contributions sur la constitution de bases de données pour la conception de systèmes d'intelligence artificielle

Afin de bénéficier de l'expertise pratique et opérationnelle des acteurs de l'IA, la CNIL a souhaité recueillir les contributions de tous les acteurs concernés sur plusieurs points structurants de l'analyse :

- **la question de la finalité (objectif), notamment pour les IA à usage général ;**
- **les méthodes de sélection, de nettoyage et de minimisation des données disponibles à l'état de l'art ;**
- **les approches visant à prendre en compte la protection des données par défaut et dès la conception ;**
- **les critères à prendre en compte si l'intérêt légitime est la base légale du traitement de collecte de base de données et du traitement de configuration (appelé parfois « entraînement ») du modèle d'intelligence artificielle.**

Ainsi, tout acteur privé ou public concerné était invité à participer à cet appel à contributions, notamment par des exemples concrets de situations rencontrées. Les contributions étaient libres, et il n'était pas nécessaire de répondre à l'ensemble des questions soulevées dans le questionnaire.

La CNIL a reçu les réponses de 9 participants à l'appel à contribution parmi lesquels :

- **5 entreprises privées ;**
- **1 institut de recherche ;**
- **1 établissement de santé ;**
- **1 syndicat professionnel de salariés ;**
- **1 particulier.**

Les entreprises privées ayant répondu à l'appel d'offre sont spécialisées dans des domaines divers, tels que l'anonymisation de données, le marketing numérique, la sécurité de l'IA, l'identité numérique et la défense.

Définir une finalité

D'une manière générale, les participants confirment la difficulté de définir une finalité pour la conception d'un système d'IA dans certains cas.

L'approche proposée qui repose sur une référence aux modes de réutilisations, aux tâches et capacités du système semble accueillie favorablement dans l'ensemble.

Par ailleurs, les participants suggèrent que :

- la finalité soit définie par une référence aux modes de réutilisation envisagées ou envisageables des données personnelles, et à titre facultatif par la capacité ou la tâche du modèle ;
- dans le cas de la recherche, le domaine de recherche soit précisé (recherche démographique, médicale, en audit algorithmique, etc.) ;
- d'élargir la définition des capacités ou tâches principales du modèle, comme par exemple la génération de contenu synthétique (par exemple texte, image, vidéo) ou l'identification d'objets / contenu ;

Choisir une base légale et traiter des données sensibles

Un participant a considéré que le consentement des personnes était la base légale la plus adaptée (si ce n'est la seule envisageable), en particulier dans le cadre de la conception d'IA génératives. Il a également soulevé que la négociation d'accords collectifs pourrait s'avérer pertinente dans ce cas.

Un participant a soulevé qu'il était difficile de réaliser la balance des intérêts nécessaire à la mobilisation de l'intérêt légitime pour l'amélioration des algorithmes utilisés dans les dispositifs médicaux à partir de données collectées lors de soins.

Les participants ont indiqué que, pour le traitement de données dites « sensibles », notamment biométriques, le consentement semble difficile à mettre en œuvre dans le cas de l'IA.

Qualifier un traitement de « recherche scientifique », au sens du RGPD

Les participants suggèrent que :

- l'utilisation dans un processus commercial ne soit pas compatible avec le statut de recherche scientifique ;
- la source du financement soit prise en compte pour apprécier s'il s'agit d'une recherche scientifique ;
- la publication du modèle, ou des techniques d'apprentissage et d'évaluation dans une revue scientifique soient des pistes à prendre en compte pour identifier un traitement de données à des fins de recherche scientifique ;
- la notion de contribution méthodologique, comme le développement d'une méthode d'IA relativement générique pour aborder une famille de problématiques, soit prise en compte pour qualifier le traitement de recherche scientifique ;
- la diffusion en source ouverte des modèles soit prise en compte, l'open source étant une composante essentielle de la recherche en IA ;
- il soit possible de rester dans la recherche scientifique sous réserve du respect de certains critères (publication, reproductibilité, etc.), par exemple pour un modèle librement accessible sous une licence en open source, si seuls les services de support sont commercialisés (matériel, logiciel, amélioration du modèle, etc.) ;
- la qualification du système comme dispositif médical dans le domaine de la santé, ou l'attribution d'un titre de propriété intellectuelle soient utilisés pour distinguer la finalité de recherche de la finalité commerciale en droit.

Certains participants ont suggéré que des bacs à sable réglementaires soient mis en place pour combler l'écart entre les exigences relatives aux modèles associés à la recherche et à des fins commerciales, afin d'encadrer la transition de la recherche à l'utilisation commerciale.

Minimiser les données

➢ **Pour la sélection des données**

Les participants ont suggéré :

- L'utilisation de l'apprentissage actif (*active learning*) après la collecte et avant l'annotation ;
- L'identification des données non pertinentes après la collecte, car il aura été possible d'évaluer leur utilité pour l'apprentissage, et tout stockage ou transfert non nécessaire pourra alors être évité ;
- De favoriser les techniques visant à réduire la précision des données, tout en tentant compte de l'impact de ces techniques sur les capacités d'apprentissage du modèle, et de réduction des biais ;
- D'étudier l'impact sur les performances obtenues suite à une suppression de variables ou une réduction du volume de données ;
- D'avoir recours aux techniques de sélection des variables pertinentes et d'ingénierie des caractéristiques (*features engineering*) de type *principal component analysis* ;
- De mesurer l'impact de la minimisation sur la mesure et la réduction des biais ;

- De privilégier la synthèse de données pour éviter une collecte (notamment pour les études a posteriori comme pour comparer des modèles ou en faire la validation).

➢ **Pour déterminer le processus de développement de l'IA le plus adapté (et les données associées)**

Les participants ont relevé plusieurs avantages et inconvénients à chacun des modèles organisationnels de développement d'un système d'IA, notamment :

- Pour le développement en interne : ce modèle permet davantage de contrôle et de personnalisation mais demande plus d'expertise et des ressources spécialisées ;
- Pour le développement externalisé : ce modèle peut s'avérer plus efficace et rentable mais peut induire une perte de contrôle ;
- Pour l'utilisation de modèles pré-entraînés ou l'apprentissage par transfert : ce modèle peut s'avérer plus rapide et demander moins de collecte, mais nécessite que le modèle de base corresponde à la tâche recherchée.

Par ailleurs, certains participants ont mis en avant que les critères les plus adaptés semblent être la performance d'apprentissage automatique, la vitesse, le coût de calcul (en investissement de capital et en temps), la maintenabilité du système (dont le débogage) et la robustesse du système.

Constituer une base de données de qualité

➢ **Pour la qualité des données**

Les participants ont recommandé :

- le recours à des examens aléatoires (notamment pour détecter la présence de données générées par l'IA), à la validation croisée des annotations (annotation en parallèle et indépendante par deux personnes, mesure de l'accord inter-annotateurs) ;
- le recours à des processus automatisés, notamment pour détecter une dérive des données, ou pour vérifier la compatibilité des données provenant de plusieurs sources.

➢ **Pour la réduction des biais**

Les participants ont recommandé le recours au rééchantillonnage des données, à des algorithmes tenant compte de l'équité, à la repondération (*re-weighting*), au réentraînement périodique, ou encore à l'entraînement contradictoire (*adversarial learning*).

Ils ont également fait part du besoin de recommandations concernant le traitement de données sensibles pour la mesure de biais.

➢ **Pour la représentativité**

Les participants ont recommandé :

- d'étudier la distribution statistique des données (moyenne, médiane, écart type) ;
- le recours à des techniques d'échantillonnage (échantillonnage aléatoire, stratifié ou systématique, sur/sous-échantillonnage), de validation de la généralisation (méthode de validation *k-fold* par exemple) ;
- le recours à la synthèse de données (bien que celles-ci ne puissent s'utiliser en gros volume car elles ne sont généralement par assez représentatives des paramètres réels) ;
- la prise en compte du besoin en représentativité dans les suites données à l'exercice des droits.

Protéger les données

➢ **Pour la conservation des données**

Les participants ont mis en avant le besoin d'adapter la durée de conservation des données au besoin de réentraînements des modèles.

- **Pour la protection des données dès la conception et par défaut**

Plusieurs recommandations ont été mises en avant par les participants :

- L'utilisation de certaines techniques et notamment de la confidentialité différentielle, le calcul multipartite, l'apprentissage fédéré ainsi que les salles blanches de données, ou « *data clean rooms* » (accompagné de mesures de gouvernance appropriées) ;
- L'utilisation de filtres sur les entrées et sorties prévus lors de la conception pour vérifier que seules les données pertinentes et appropriées sont traitées ;
- L'utilisation de tables de correspondance pour la pseudonymisation des données de contact (nom, prénom, adresse e-mail) liées aux données biométriques ;
- La mise en œuvre d'une gouvernance dédiée à la gestion des données personnelles et des données biométriques et sa documentation ;
- La mise en place d'un mécanisme permettant de tracer les données et le consentement qui en a autorisé la collecte, tant pour faciliter le respect des droits prévus par le RGPD que les droits de propriété intellectuelle, y compris à l'aide d'un tatouage numérique (*watermarking*) ;
- L'anonymisation dans les cas où cela est possible, bien que cela ne soit pas le cas dans certains domaines (comme la génomique ou les réponses fournies à un chatbot) ;
- Une mise en œuvre de l'anonymisation ou de la pseudonymisation dès que possible, et notamment par le prestataire réalisant la collecte lorsque cela est réalisable ;
- Un traitement local des données pour l'apprentissage des modèles ou une extraction limitée à des informations pseudonymes ou anonymes issues d'un traitement algorithmique, comme les représentations des données dans l'espace latent, notamment sur les sites web ;

Un participant a toutefois soulevé qu'il pouvait être difficile pour un organisme de vérifier si les procédés tels que la confidentialité différentielle pouvaient être qualifiés d'anonymisation, tant en raison du risque juridique existant en cas d'erreur d'appréciation que d'un manque de compétence technique (en particulier lorsque ces procédés sont proposés par un prestataire).

- **Pour la sécurité**

Les participants ont suggéré :

- L'utilisation de certaines techniques : les environnements d'exécution fiables ou encore les enclaves sécurisées (bien qu'un risque existe puisque les acteurs fournissant ces services pourraient recouper les données qui leur sont fournies et que ces dispositifs aient un coût parfois élevé) ;
- Le recours à des tests de pénétration (attaques fictives) pour valider la sécurité ;
- Le chiffrement des données au repos et en transit ;
- D'évaluer régulièrement la sécurité des modèles contre les attaques (attaques par exfiltration, attaque d'extraction de modèle).