IA : Tenir compte de la protection des données dans la conception du système

08 avril 2024

Pour assurer le développement d'un système d'IA respectueux de la protection des données, il est nécessaire de mener une réflexion préalable lors de la conception du système. Cette fiche en détaille les étapes.

Lors de la réflexion autour des choix de conception d'un système d'IA, les principes de protection des données, et en particulier le principe de minimisation, doivent être respectés. Cette démarche s'opère à cinq niveaux. Un responsable de traitement doit ainsi s'interroger sur :

- **l'objectif du système** qu'il souhaite développer ;
- **la méthode à employer** qui aura une incidence sur les caractéristiques de la base de données ;
- **les sources de données mobilisées** (voir la fiche sur la conformité du traitement à la loi, sur les sources ouvertes, sur les tiers, etc.) et ;
- **parmi ces sources, la sélection des données strictement nécessaires**, au regard de l'utilité des données et de l'impact potentiel que leur collecte fait peser sur les droits et libertés des personnes concernées ;
- **la validité des choix** précédemment opérés. Cette validation peut pendre différentes formes (non exclusives) telles que la réalisation d'une **étude pilote** ou l'avis d'un **comité éthique**.

L'objectif du système

L'objectif de cette étape est de concevoir, sur la base de la finalité identifiée (voir Fiche n° 2), un système conforme à un cahier des charges, tout en limitant les conséquences potentielles pour les personnes concernées.

En précisant l'emploi du système d'IA en phase de déploiement (que celui-ci soit mis en œuvre directement par le fournisseur ou par un tiers), le fournisseur du système doit déterminer :

- le type de résultat / sortie attendu ;
- des indicateurs de la performance acceptable de la solution, qu'il s'agisse d'indicateurs quantitatifs (F1-score, erreur quadratique moyenne, temps de calcul) ou qualitatifs (provenant de retours humains par exemple) ;
- le contexte d'utilisation du système permettant d'identifier les informations prioritaires pour son usage opérationnel ;
- les contextes d'utilisation exclus et les informations non pertinentes pour le ou les cas d'usage principaux envisagés du système.

Certaines techniques d'IA peuvent permettre de réaliser des tâches complexes qui dépassent les objectifs initiaux des fournisseurs. En définissant précisément la fonctionnalité attendue, il est ainsi possible d'éviter des risques de sur-collecte.

Exemple : pour réaliser l'apprentissage d'un système d'IA visant à permettre le comptage des personnes qui se tiennent debout dans un tramway à partir d'images de caméras de vidéoprotection, les systèmes suivants sont techniquement envisageables :

- un réseau de neurones détectant la présence de personnes dans un wagon, sans analyse de la posture, intégré dans un algorithme réalisant un décompte des personnes debout (le nombre de personnes debout pouvant être déduites grâce au nombre de places assises) ;
- un réseau de neurones réalisant une analyse de la posture des personnes dans un wagon intégré dans un algorithme réalisant un décompte des personnes qui se tiennent debout.

Le premier réseau pourrait fournir moins d'informations (notamment le décompte de personnes debout). Toutefois, si l'estimation donnée est suffisante pour le cas d'usage envisagé, en particulier pour le calcul de statistiques d'occupation, il est alors préférable d'avoir recours à ce modèle. En effet, celui-ci nécessitera une quantité de données plus faible pour son apprentissage tout en permettant de remplir l'objectif poursuivi, alors que le second demande une collecte et une annotation de données spécifiques et de plus grande ampleur. Le principe de minimisation converge alors avec la réduction des coûts de conception du système, sans préjudice pour la précision du système.

La méthode à employer

Bien souvent, une même tâche peut être réalisée par différentes techniques. Toutefois, toutes ne sont pas équivalentes, car elles n'impliquent pas de recourir aux mêmes données et pas nécessairement les mêmes quantités. Elles peuvent ne pas permettre d'atteindre le même niveau de performances, présenter des enjeux plus ou moins importants en termes d'explicabilité, être soumises à différentes contraintes opérationnelles (comme le coût de calcul). Tout en prenant en compte ces enjeux, **le fournisseur du système doit sélectionner la technique la plus respectueuse des droits et libertés des personnes afin de respecter le principe de minimisation des données en tenant compte de l'objectif recherché**. Autrement dit, si une technique remplit la même fonction/permet d'atteindre le même résultat avec moins de données personnelles, elle doit être préférée.

En particulier, les méthodes d'apprentissage machine nécessitent l'utilisation d'un très grand nombre de données. Afin d'assurer le respect du principe de proportionnalité et de minimisation des données, le recours à ces solutions techniques doit donc être justifié. S'il existe une méthode n'utilisant pas l'apprentissage machine et que celle-ci permet de remplir les objectifs poursuivis, celle-ci doit être privilégiée. Le recours à l'apprentissage profond est à justifier et ne doit donc pas être systématique.

Exemples :
Pour assurer la sécurité des employés, un fournisseur veut délimiter une zone dangereuse à ne pas franchir dans son entrepôt. Il souhaite une solution qui permette de détecter la présence d'une personne dans cette zone et de déclencher un signal sonore pour avertir la personne du danger. Le recours à un détecteur de mouvement infrarouge doit être privilégié en lieu et place d'une caméra augmentée, car cette

solution ne collecte pas l'image des personnes, conformément au principe de minimisation des données et de protection des données dès la conception.

L'analyse sémantique du contenu d'un texte pourrait être réalisée par un réseau neuronal entraîné sur une base de données textuelles annotées, par une méthode ensembliste tel qu'une forêt d'arbres décisionnels ou encore par un algorithme non supervisé, tel qu'un algorithme de partitionnement des données (« *clustering* »).

Au stade de l'entraînement du modèle, il y a aussi lieu de tenir compte de l'incertitude éventuelles sur les performances de telle ou telle architecture : le respect du principe de minimisation s'apprécie en fonction des connaissances scientifiques disponibles.

Selon les avancées dans le domaine concerné, cette réflexion doit reposer sur plusieurs facteurs pour chacune des architectures considérées. Cette analyse technique peut se faire grâce à :

- **un état de l'art**, au moyen, par exemple :
    - d'une étude de la littérature scientifique (recensement et étude des publications académiques ou privées, conférences spécialisées, etc.) ;
    - d'une enquête auprès des professionnels du domaine : la démarche d'ouverture du code informatique (y compris par placement sous licence libre) de certains acteurs du secteur contribue à rendre possible une comparaison des techniques ;
    - de la sollicitation de la communauté spécialisée (compétitions en ligne, forums en ligne, conférences et rencontres dédiées, etc.) ;

- une comparaison des résultats obtenus après l'implémentation de plusieurs architectures sous la forme de « preuves de concept » ;
- une comparaison des résultats obtenus par l'utilisation d'un modèle existant et pré-entraîné (pouvant nécessiter éventuellement un ajustement, ou *fine-tuning*) et d'un modèle développé par le fournisseur.

Si le choix du modèle d'IA et des algorithmes utilisés peuvent permettre de limiter la collecte de données, d'autres choix de conception sont à prendre en compte, notamment dans une optique de protection des données dès la conception. Le choix du protocole d'apprentissage utilisé notamment, peut permettre de limiter l'accès aux données aux seules personnes habilitées, ou encore de ne donner accès qu'à des données chiffrées. Deux techniques, applicables dans certaines situations, sont particulièrement intéressantes :

- Les protocoles d'apprentissage décentralisés, tels que **l'apprentissage fédéré**, technique à laquelle un article LINC a été dédié. Cette technique permet d'entraîner un modèle d'IA depuis plusieurs bases, et ainsi à chacun des acteurs de la chaîne de conserver la main sur ses données. Cette technique possède toutefois certains risques, concernant la sécurité des bases décentralisées, ainsi que sur la confiance entre les acteurs parmi lesquels un acteur malicieux pourrait conduire une attaque par empoisonnement des données par exemple.
- Les **ressources offertes par la cryptographie**. Les progrès scientifiques récents dans le domaine de la cryptographie peuvent permettre d'obtenir des garanties fortes pour la protection des données. En fonction des cas d'usage, il pourra par exemple être pertinent d'explorer les possibilités offertes par le calcul multipartite sécurisé (« *secure multi-party computation* »), ou encore le chiffrement homomorphe (« *homomorphic encryption* »). Les techniques utilisées dans ce domaine permettent d'entraîner un modèle d'IA sur des données qui restent chiffrées tout au long de l'apprentissage. Elles demeurent toutefois limitées en ce qu'elles ne peuvent pas être appliquées à tous types de modèles et

en raison du temps de calcul supplémentaire qu'elles induisent. Par ailleurs, certaines d'entre elles, comme le chiffrement homomorphe pour l'entraînement de réseaux de neurones, font encore l'objet d'études. Les évolutions techniques étant fréquentes dans ce domaine, il est conseillé de maintenir une veille active sur ce sujet.

Cette liste de mesures n'est pas exhaustive, des mesures supplémentaires pourraient être citées comme le recours à un environnement d'exécution sécurisé (ou « *trusted execution environment* »), à la confidentialité différentielle appliquée lors de l'apprentissage ou encore des mesures permettant d'anticiper le désapprentissage machine. Plus généralement, en raison de l'évolution rapide de la technique, il est recommandé de mener une veille technologique sur les pratiques protectrices de la vie privée applicables lors du développement de systèmes d'IA.

La sélection des données strictement nécessaires

Le principe

Le principe de minimisation prévoit que les données personnelles doivent être **adéquates, pertinentes et limitées** à ce qui est nécessaire au regard des finalités pour lesquelles elles sont traitées. Une attention particulière doit être apportée à la nature des données et ce principe doit être appliqué de manière particulièrement rigoureuse lorsque les données traitées sont sensibles (au sens de l'article 9 du RGPD).

En pratique

Le principe de minimisation ne signifie pas qu'il est interdit d'entraîner un algorithme avec des volumes très importants de données : il implique d'avoir une réflexion en amont de l'entraînement pour ne pas recourir à des données personnelles qui ne seraient pas utiles au développement du système. Pour identifier les données personnelles nécessaires au développement d'un système d'IA, il convient de prendre en compte quatre dimensions :

- **Le volume** : nombre de personnes concernées, profondeur historique, précision des données, répartition des données selon les situations et populations et couvrir, etc. Il pourra être justifié, par exemple, par les capacités de calcul limitées des serveurs utilisés pour l'apprentissage, les besoins en termes de représentativité du jeu de données, les pratiques communément admises par la communauté scientifique, une comparaison des résultats obtenus en faisant varier le volume de données, une analyse statistique démontrant qu'un volume minimum de données est nécessaire pour atteindre des résultats significatifs, etc. ;
- **Les catégories** : âge, sexe, image du visage, activité sur un réseau social, etc. La présence de données sensibles ou de données à caractère hautement personnel doit être examinée et justifiée (voir Fiche n° 3). Cette analyse peut reposer sur la nécessité d'entraîner le modèle sur des données contrefactuelles (susceptibles de donner lieu à des faux positifs en pratique), sur une étude de l'utilité des catégories de données concernées (voir encadré plus bas), etc. Parmi ces catégories de données, il convient de privilégier le format le moins intrusif sans perte d'information pour l'objectif poursuivi, par exemple l'âge ou tranche d'âge plutôt qu'une date de naissance complète ;

- **La typologie** : données réelles, de synthèse, augmentées, issues de simulation, des données anonymisées ou pseudonymisées, etc. ;
- **Les sources** : comme précisé dans la Fiche n° 3, recensement des sources de données auxquelles il est envisagé de recourir, qu'il s'agisse d'une collecte initiale ou d'une réutilisation (données disponibles en source ouverte, collectées précédemment par le fournisseur ou encore auprès de fournisseurs de données).

Bien que la sélection des données soit une phase généralement nécessaire afin de concevoir un système d'IA sur la base de données de qualité, dans certains cas et à titre subsidiaire, il peut être possible de traiter un ensemble de données de manière indiscriminée. La nécessité devra alors en être justifiée.

Outre la prise en compte de ces dimensions d'ordre technique, **une attention particulière devra être portée à la nature des données au sens du RGPD**, et plus particulièrement s'il s'agit de données sensibles ou hautement personnelles.

À noter :

Les questions relatives à la **distribution** et à la **représentativité des données** doivent également être traitées lors de cette étape. Celles-ci sont en effet essentielles pour **limiter au maximum les risques de biais de discrimination.**

Au sein de cette question se pose notamment celle de l'inclusion de données de type « vrais négatifs » dans la base de données d'apprentissage (notamment pour le test et la validation en vue de vérifier l'absence de certains effets de bord ou d'apprentissage).

Ces questions étant particulièrement importantes, une fiche dédiée sera prochainement publiée.

La validité des choix de conception

À l'issue des trois étapes précédentes, les choix de conception sont théoriquement validés et la collecte des données peut commencer. Afin de valider quantitativement et qualitativement les choix de conceptions, plusieurs mesures sont recommandées à titre de bonne pratique.

Mener une étude pilote

L'objectif du pilote est de s'assurer que les choix d'ordre technique et ceux relatifs aux types de données identifiés sont bien pertinents. Pour ce faire, une expérimentation à petite échelle peut est réalisée. Des données fictives, synthétiques, anonymisées ou à défaut des données personnelles collectées conformément au RGPD peuvent être utilisées.

Exemples :
L'utilisation de données issues de réseaux sociaux sur les pages personnelles de personnes ayant consenti à la collecte de leurs données.

Ce type d'expérimentation n'offre pas toujours une vision représentative de l'activité rencontrée sur les réseaux sociaux, mais elle peut être adaptée à certains cas d'usage comme l'identification de contenus haineux ou l'étude du ciblage publicitaire sur ces

réseaux. Cette pratique est bénéfique car elle offre un niveau de transparence largement supérieur à certaines pratiques de moissonnage (« *web scraping* »).

La conception d'un système de recommandation de films

Un organisme peut collecter auprès d'utilisateurs volontaires la liste de films visionnés sur une semaine et ceux visionnés dans les jours qui ont suivi, soit par données déclaratives, soit en collectant leur historique de visionnage sur des sites dédiés. Il peut mener son étude pilote sur les données ainsi collectées en anonymisant les identifiants de chaque utilisateur.

Interroger un comité éthique

L'association d'un comité éthique au développement de systèmes d'IA est une bonne pratique pour garantir que les enjeux en matière d'éthique et de protection des droits et libertés des personnes soient pris en compte en amont.

Le comité éthique peut avoir plusieurs missions :

- la formulation d'avis sur tout ou partie des projets, outils, produits, etc. de l'organisme susceptibles de problématiques éthiques, qui lui seraient soumis ;
- l'animation d'une réflexion et l'élaboration d'une doctrine interne sur les aspects éthiques du développement de systèmes d'IA par l'organisme (par exemple, quelles conditions pour le recours à la sous-traitance) ;
- la mise au jour d'attitudes collectives et individuelles et la recommandation de certains principes, comportements ou pratiques.

La constitution et le rôle de ce comité peuvent varier selon les situations, mais plusieurs bonnes pratiques sont recommandées. Le comité devrait :

- **être pluridisciplinaire** : les profils des membres du comité – employés de l'organisme et/ou personnes externes – doivent être diversifiés. Les personnes amenées à y siéger contribuent aux missions du comité et peuvent mettre à jour des problématiques que les équipes de développement n'avaient pas envisagées. Une bonne pratique est d'attribuer certains sièges du comité aux employés de l'organisme qui l'occuperont chacun leur tour. En outre, la diversité des membres du comité du point de vue du sexe, de l'âge et des origines ethniques et culturelles est vivement encouragée ;
- **être indépendant** : les avis rendus par le comité peuvent avoir des implications importantes, par exemple pour la direction commerciale d'une entreprise et ainsi favoriser ou défavoriser certains de ses projets. Ainsi, les personnes siégeant au comité ne doivent pas être motivées par un éventuel gain (qu'il soit financier ou d'un autre ordre) à tirer par la décision rendue. De même, lorsque des employés siègent au comité, les décisions rendues ne doivent pas avoir de conséquences pour eux ;
- **avoir un rôle clairement défini** : afin de garantir l'intégration systématique du comité, une procédure doit être fixée afin de déterminer les conditions dans lesquelles le comité se réunit et doit être associé. Selon les situations, le comité peut être simplement consultatif ou adopter des avis contraignants : les deux approches présentent des avantages et des inconvénients. Si le comité émet des avis contraignants, son insertion dans la gouvernance d'entreprise doit être particulièrement bien définie au regard des statuts de l'organisme, pour éviter son instrumentalisation. Si le comité est consultatif, son impact doit être garanti, notamment en garantissant sa saisine obligatoire selon des critères précis et la transparence large de ses avis, a minima au sein de l'organisme et éventuellement

d'autres mesures comme l'obligation pour le porteur de projet de répondre par écrit aux remarques du comité ;
- **être averti** : le comité est encouragé à s'informer, documenter ses avis et partager ses connaissances. Les risques que comporte l'utilisation de l'IA évoluent avec le développement technique et les nouveaux usages dans ce domaine, et il est nécessaire de s'informer, notamment grâce à la littérature académique et aux publications des entités compétentes dans ce domaine (comme le Défenseur des Droits, ou le Comité national pilote d'éthique du numérique). La diffusion des connaissances acquises permettra d'appuyer les avis rendus et de répandre certaines bonnes pratiques.

Dans le cas du développement d'un système d'IA, l'avis du comité éthique pourrait être sollicité sur plusieurs questions :

- Les données utilisées pour le développement respectent-elles les critères éthiques de l'organisme ?
- Les usages opérationnels prévus pour le système d'IA pourraient-ils entraîner des conséquences individuelles ou sociétales graves ? Ces conséquences peuvent-elles être évitées ? Ces usages opérationnels peuvent-ils être exclus ?
- Les potentielles utilisations détournées du système d'IA (qu'elles soient volontaires ou accidentelles, en particulier pour les modèles diffusés en source ouverte) pourraient-elles entraîner des conséquences graves pour les personnes ou pour la société ? Quelles mesures permettraient de les éviter ?
- Les choix techniques sont-ils suffisamment maîtrisés par l'organisme (dans le cas du recours à des approches radicalement nouvelles) ?
- Les mesures de transparence, pour l'exercice des droits des personnes ou leur permettant d'exercer un éventuel recours sont-elles suffisantes ?
- Les discriminations que peuvent entraîner l'utilisation du système sont-elles répertoriées et les moyens nécessaires ont-ils été mis en œuvre afin d'éviter leur survenue ?
- L'organisme est-il organisé de manière à prévenir les risques dès la conception (que ce soit en matière de discriminations, de protection des données, de protection du droit d'auteur, de sécurité informatique, etc.) ?

En fonction de la taille des organismes et de la façon dont ils sont structurés, il n'est pas toujours envisageable de constituer un comité éthique. Néanmoins, il est essentiel que de telles réflexions puissent être menées pour accompagner le développement de systèmes d'IA. La nomination d'un « référent éthique » peut être une alternative permettant la prise en compte de ces questionnements.

< Fiche précédente : Réaliser une analyse d'impact si nécessaire

Sommaire

Fiche suivante : Tenir compte de la protection des données dans la collecte et la gestion des données >