IA : Assurer que le traitement est licite - En cas de réutilisation des données, effectuer les tests et vérifications nécessaires

08 avril 2024

En cas de réutilisation de données, le responsable du traitement est tenu d'effectuer certaines vérifications supplémentaires afin de garantir que le traitement de données est autorisé par la loi. La CNIL vous aide à déterminer vos obligations en fonction des modalités de collecte et de la source des données

Le principe

Dans certains cas, en fonction des modalités de collecte et de la source des données utilisées pour la constitution de la base de données d'apprentissage, le responsable du traitement est tenu d'effectuer certaines vérifications afin de garantir que le traitement de données est autorisé par la loi. **Ces vérifications s'ajoutent à l'identification de la base légale du traitement de données.**

En pratique

Le fournisseur réutilise les données qu'il a lui-même collectées initialement pour une autre finalité

Un responsable de traitement peut vouloir réutiliser les données qu'il a collectées pour une finalité initiale (par exemple, dans le cadre de la fourniture d'un service à des particuliers) afin de constituer une base de données à des fins d'apprentissage d'un système d'IA.

Dans ce cas, **il doit déterminer si ce traitement ultérieur est compatible avec la finalité pour laquelle les données ont été initialement collectées**, lorsque le traitement ne s'appuie pas sur le consentement de la personne concernée ou sur le droit de l'Union ou le droit d'un État membre.

L'obligation d'effectuer ce « test de compatibilité » s'applique aux traitements ultérieurs de données, (au sens de l'article 6.4 du RGPD), c'est-à-dire ceux :

- qui n'ont pas été prévus ni portés à la connaissance des personnes concernées lors de la collecte des données ;
- qui sont effectués par un même responsable de traitement qui décide de réutiliser des données pour une finalité distincte de celle pour laquelle elles ont été collectées, y compris quand il s'agit de les publier sur

Internet ou de les partager avec des tiers à des fins de réutilisation pour une autre finalité.

À noter :

Aucun test de compatibilité n'est requis pour les finalités prévues et portées à la connaissance des personnes concernées dès la collecte dans le respect du principe de transparence, y compris lorsque certaines d'entre elles peuvent paraître secondaires ou accessoires. Par exemple, le partage de données par un responsable de traitement avec son sous-traitant pour l'amélioration de la performance de son algorithme ne nécessite pas d'effectuer un test de compatibilité, si cette finalité était prévue et portée à la connaissance de la personne concernée (sous réserve de respecter les conditions de légalité pour cette finalité d'amélioration de l'algorithme).

Pour réaliser ce « test de compatibilité » il doit notamment prendre en compte :

- l'existence d'un lien entre la finalité initiale et la finalité du traitement ultérieur envisagé ;
- le contexte dans lequel les données personnelles ont été collectées, en particulier les attentes raisonnables des personnes concernées, en fonction de la relation entre les personnes concernées et le responsable du traitement ;
- le type et la nature des données, en particulier en fonction de leur sensibilité (données biométriques, de géolocalisation, concernant des mineurs, etc.) ;
- les éventuelles conséquences du traitement ultérieur envisagé pour les personnes concernées ;
- l'existence de garanties appropriées (telles que le chiffrement ou la pseudonymisation).

Exemples :

Le fournisseur d'un éditeur de texte grand public lance une fonctionnalité d'IA générative permettant de compléter certaines phrases ou certains paragraphes (auto-saisie). Quelques temps après le déploiement de cette fonctionnalité, il souhaite réutiliser les corrections manuelles apportées par les utilisateurs au contenu des textes ainsi générés, afin de proposer à chaque utilisateur de disposer d'une version personnalisée de son service de recommandation (par exemple pour mieux comprendre et anticiper sa manière d'écrire) sur la base de leurs données respectives. Une plateforme de streaming vidéo grand public envisage désormais de réutiliser l'historique et les listes de lecture qu'elle a enregistrés dans le cadre de la fourniture du service pour proposer à chaque utilisateur de disposer d'une version personnalisée de son service de recommandation (par exemple pour mieux anticiper et comprendre ses préférences) sur la base de leurs données respectives.

Dans ces deux cas, la nouvelle finalité pourra être considérée comme compatible avec la finalité initiale de la fourniture du service, à condition que les garanties mises en œuvre soient suffisantes (par exemple, grâce à la possibilité de s'opposer à cette réutilisation, sans avoir à fournir de motif) sur la base de leurs données respectives.

Lorsque la réutilisation des données poursuit des fins statistiques ou de recherche scientifique, le traitement est présumé compatible avec la finalité initiale s'il respecte le RGPD et s'il n'est pas utilisé pour prendre des décisions à l'égard des personnes concernées. La réalisation du « test de compatibilité » n'est donc pas nécessaire.

Pour poursuivre une finalité statistique au sens du RGPD, le traitement ne doit tendre qu'à la production de données agrégées pour elles-mêmes : le traitement doit avoir pour unique objet le calcul des données, leur affichage ou publication, leur éventuel partage ou communication (et non à la prise de décisions ultérieures, individuelles ou collectives). Les résultats statistiques ainsi obtenus doivent constituer des données agrégées et anonymes au sens de la réglementation sur la protection des données. Le recours à des techniques statistiques d'apprentissage automatique ne suffit pas à considérer qu'il s'agit de traitements « à des fins statistiques », dans la mesure où la finalité du traitement de données n'est pas de produire des données agrégées pour elles-mêmes. Le recours à cette technique est davantage un moyen mis en œuvre pour l'apprentissage du modèle.

La notion de « recherche scientifique » est entendue largement dans le RGPD. En synthèse, la recherche a pour objet de produire des connaissances nouvelles dans tous les domaines dans lesquels la méthode scientifique est applicable. Tout traitement de données à des fins de recherche scientifique doit être soumis à des garanties appropriées pour les droits et libertés de la personne concernée, telles que l'anonymisation ou la pseudonymisation (mentionnées à l'article 89 du RGPD).

En savoir plus :

- La recherche scientifique (hors santé)
- La réutilisation de données publiquement accessibles à des fins de recherche scientifique (hors santé), extrait du guide soumis à consultation publique

À noter :
Même lorsque le traitement ultérieur est compatible, une base légale valable doit toujours être identifiée et les personnes informées, notamment pour pouvoir exercer leurs droits..

Focus : à quelles conditions peut-on réutiliser un jeu de données initialement constitué à des fins de recherche scientifique ?

Le RGPD facilite la réutilisation de données à des fins de recherche scientifique : cette réutilisation est jugée compatible avec la finalité initiale du traitement et certaines dérogations (notamment aux droits des personnes) sont possibles.

En revanche, lorsqu'un responsable du traitement a traité des données à des fins de recherche scientifique et qu'il entend les réutiliser à d'autres fins (pour son propre compte ou pour les transmettre à un tiers), il doit respecter certaines conditions. À cet égard, **la réutilisation d'un jeu de données sera possible** :

- **si les données ont été préalablement anonymisées, ou**
- **si la réutilisation est compatible avec la finalité pour laquelle le responsable du traitement a collecté les données** (selon le « test de compatibilité » détaillé ci-dessus) **et que le nouveau traitement est mis en œuvre dans le respect du RGPD** (information des personnes au sujet de cette nouvelle finalité, identification d'une base légale, etc.). Les dérogations permises par le RGPD pour la recherche scientifique ne seront plus mobilisables.

En cas de transmission des données à des tiers, la compatibilité des réutilisations ultérieures avec la finalité de recherche pourra être garantie notamment par une licence de réutilisation.

Le fournisseur réutilise une base de données publiquement accessibles

Des bases de données contenant des données personnelles peuvent être librement mises à disposition sur Internet en dehors du cadre légal relatif à l'ouverture des données (« *open data* »). Le plus souvent, il s'agit de données qui étaient déjà publiquement accessibles et qui constituent une base de données ou un corpus diffusé sur le site web d'une université ou d'une plateforme dédiée au partage de bases de données, pour faciliter leur réutilisation.

Le contrôle du caractère licite de la mise en ligne de la base de données relève en premier lieu du responsable du traitement qui opère cette mise en ligne (le cas échéant en s'assurant qu'il s'agit d'un traitement ultérieur compatible s'il n'a pas initialement collecté les données dans ce but). Cependant, afin de pouvoir se prévaloir d'une base légale au titre du RGPD, le responsable du traitement qui réutilise les données doit s'assurer qu'il n'est pas en train de réutiliser une base de données dont la constitution était manifestement illicite (par exemple, provenant d'une fuite de données).

Le réutilisateur ne peut pas réutiliser une base de données constituée ou mise en ligne dont il ne peut ignorer qu'elle ne respecte pas le RGPD (article 5.1.a du RGPD) ou d'autres règles, telles que celles interdisant les atteintes à la sécurité des systèmes d'information ou les atteintes à des droits de propriété intellectuelle.
En outre, la personne qui télécharge ou réutilise une base de données manifestement illégale risque de se rendre coupable du délit de recel (article 321-1 du code pénal).

Si la possibilité de réutiliser une base de données librement mise à disposition sur Internet n'est pas nécessairement subordonnée à des vérifications approfondies sur le respect de l'ensemble des règles du RGPD ou d'autres règles juridiques applicables (droit d'auteur, données couvertes par le secret des affaires, etc.), vérifications qui relèvent en premier lieu de l'organisme qui met en ligne les données, un organisme ne peut réutiliser une base de données qui serait manifestement illicite.

Cette illicéité manifeste doit s'apprécier au cas par cas. À ce titre, la CNIL recommande aux réutilisateurs de s'assurer :

- **Que la description de la base de données mentionne leur source.**

Exemple : une base de données dont la description expliquerait qu'elle a été constituée à partir de publications sur un réseau social professionnel nommément désigné.
À l'inverse, si une base de données contenant des images de vidéosurveillance ne précise pas la source, une telle base ne devrait pas être réutilisée avant d'avoir obtenu davantage de précisions permettant de lever les doutes quant à la conformité de sa constitution et de sa diffusion ;

- Que la constitution ou la diffusion de la base de données **ne résulte pas manifestement d'un crime ou d'un délit ou a fait l'objet d'une condamnation ou d'une sanction publique** de la part d'une autorité compétente qui a impliqué une suppression ou l'interdiction d'exploiter ultérieurement les données ;

Exemples : une entreprise souhaite constituer une base de données pour le développement d'un système d'IA de recommandation qu'il entend utiliser auprès de ses consommateurs. S'il acquiert pour cela une base de données sur le dark web

provenant, par exemple, d'une atteinte à un système de traitement automatisé punie par la loi (au sens de l'article 323-1 du code pénal), il ne saurait en ignorer l'origine délictuelle. Dans ce cas, le caractère illicite de la base de données serait alors manifeste.
Il en irait de même pour une entreprise souhaitant réutiliser une base de données pour laquelle une décision de justice a retenu une atteinte à un droit de propriété intellectuelle comme celui, particulier, des producteurs de bases de données (au sens de l'article L. 342-1 du code de la propriété intellectuelle) ;

- Qu'il n'y ait **pas de doutes flagrants sur le fait que** la base de données est licite (notamment que le traitement source ne soit pas manifestement dépourvu de base légale lorsque les données sont tellement intrusives qu'elles ne sauraient être traitées sans le consentement des personnes), en s'assurant en particulier que les conditions de collecte des données soient suffisamment documentées ;

Exemples :
Sur une plateforme d'hébergement de bases de données, une entreprise repère un ensemble compilant les trajets domicile-travail de milliers de personnes. Sa description explique qu'il s'agit de données de géolocalisation précises, non anonymes, sans en détailler la source. Dans cette hypothèse, elle ne saurait ignorer qu'il existe un doute sérieux quant à la licéité de la diffusion d'une telle base de données sans le consentement des personnes.

À l'inverse, il serait envisageable de constituer une base de données à partir d'une base de données dont la description ne laisse pas de doute flagrant quant à sa licéité. Par exemple, une base de données pseudonymisées, initialement rendues publiques par les personnes concernées sur un site web identifié et qui ne contiendrait pas de données sensibles.

Il en irait de même pour la réutilisation d'une base de données agrégées que le diffuseur présenterait comme anonymes. Par exemple, un organisme qui souhaite constituer une base de données pour entraîner un système d'IA destiné à prévoir l'impact socio-économique du vieillissement d'une population pourrait réutiliser des bases de données anonymes agrégées contenant notamment des informations démographiques (nombre de personnes actives, âge des personnes, taux de fécondité ou encore taux de dépendance des personnes âgées).

- Que la base de données ne contient **pas de données sensibles** (données de santé ou révélant des opinions politiques par exemple) **ou de données d'infraction** (au sens des articles 9 et 10 du RGPD), ou, si elle en contient, il est recommandé de mener des vérifications supplémentaires pour s'assurer que ce traitement était licite (il s'agirait principalement pour les données sensibles de s'assurer du recueil d'un consentement explicite des personnes concernées, ou que les données ont été manifestement rendues publiques par ces dernières comme cela est précisé ci-dessous et pour les données relatives à des infractions qu'une telle utilisation est rendue possible par la loi informatique et libertés).

Exemple : sur un forum en ligne, un chercheur découvre une base de données non anonymes qui contiendrait, selon sa description, les parcours de soin d'une centaine de patients atteints d'une pathologie particulière et qui proviendraient d'hôpitaux français. Dans ce cas, le chercheur devrait sérieusement douter que la diffusion de ce jeu de données soit licite compte tenu de l'encadrement des données de santé prévu par le RGPD et la loi « informatique et libertés ».

Ces vérifications préalables pourraient utilement figurer dans l'analyse d'impact relative à la protection des données (AIPD).

Certains manquements commis par le responsable des traitements de constitution et de diffusion d'une base de données n'impactent pas systématiquement et irrémédiablement la licéité des traitements mis en œuvre par le réutilisateur. Ainsi, un réutilisateur peut utiliser une base de données dont les illicéités sont mineures, à condition que la réutilisation satisfasse les exigences du RGPD.

Exemple : la fourniture de mentions d'informations incomplètes lors de la constitution ou de la diffusion de la base de données, ou un défaut de documentation adaptée de la conformité de ces traitements (qu'il est nécessaire de vérifier avec le diffuseur ou l'éditeur de la base de données).

Le fournisseur réutilise une base de données acquise auprès d'un tiers (courtiers en données, etc.)

Certains fournisseurs souhaitent constituer une base de données d'apprentissage à partir de bases de données détenues par des tiers.

Pour le tiers qui partage des données personnelles, cela implique de s'assurer de la licéité de cette transmission

- **Cas n°1 : les données ont précisément été collectées en vue d'être partagées à des fins de constitution d'une base de données pour l'apprentissage de système d'IA**

Le tiers devra **s'assurer de la conformité du traitement de transmission des données au regard du RGPD** (définition d'une finalité explicite et légitime, exigence d'une base légale, information des personnes et gestion de l'exercice de leurs droits, etc.) dont il assume la responsabilité.

- **Cas n°2 : le tiers n'a pas initialement collecté les données pour cette finalité**

Lorsque le tiers a initialement collecté les données pour d'autres finalités (par exemple dans le cadre de la fourniture d'un service aux personnes concernées), il lui **appartient de s'assurer que la transmission de ces données poursuit une finalité compatible avec celle(s) ayant justifié leur collecte**. Il devra donc réaliser un « test de compatibilité ».

À noter que le détenteur initial d'une base de données autorise parfois son utilisation dans le cadre d'un contrat de licence qui en prévoit les termes et les conditions (notamment au titre du droit de la propriété intellectuelle). Ce contrat de licence peut par exemple encadrer cette compatibilité en limitant les réutilisations possibles.

Pour le réutilisateur, cela implique le plus souvent une série de vérifications des traitements du responsable de traitement initial

Le responsable du traitement **doit s'assurer qu'il n'est pas en train de réutiliser une base de données dont la constitution ou le partage était manifestement illicite** (par exemple, en l'absence d'indication quant à sa source, en cas de doute flagrant sur sa licéité, en particulier dans le cas de traitement de données sensibles, etc.). Cela résulte du principe général de licéité des traitements de l'article 5.1.a du RGPD, outre le risque de se rendre coupable du délit de recel (article 321-1 du code pénal). Cela implique pour le responsable de traitement d'effectuer a minima les mêmes vérifications que celles énoncées dans la partie ci-dessus.

Le réutilisateur d'une base de données transmise de gré à gré par un tiers pourra d'autant moins ignorer qu'elle est constituée ou partagée en méconnaissance du RGPD ou de règles plus générales (telles que celles interdisant les atteintes à la sécurité des systèmes d'information ou des atteintes à des droits de propriété intellectuelle) que sa relation avec ce tiers lui permet de lever les doutes qu'il pourrait avoir.

La conclusion d'un accord entre le détenteur initial des données et le réutilisateur est ainsi recommandée afin de permettre à ce dernier de s'assurer de la licéité de ses propres traitements, quand bien même elle ne serait pas explicitement exigée par le RGPD.

À cet égard, la CNIL recommande de fournir un certain nombre d'indications dans le contrat telles que :

- la source, le contexte de la collecte des données, la base légale du traitement et l'analyse d'impact relative la protection des données (voir notamment la fiche n° 5 sur la réalisation d'une AIPD) si nécessaire, afin d'écarter les risques d'avoir une base de données illicite ;
- les mentions d'information des personnes portées à la connaissance des personnes (en particulier s'agissant de la finalité et des destinataires) ;
- d'éventuelles garanties quant à la licéité de ce partage de données par le détenteur initial des données (par exemple : sur la compatibilité de la finalité, sur la licéité du partage, etc.).

La CNIL fournit un modèle de fiche descriptive du jeu de données qui peut utilement être utilisé à cette fin.

À noter : si le réutilisateur souhaite fonder son traitement sur un consentement recueilli par un tiers, il doit être en mesure d'apporter la preuve qu'un consentement valide a bien été recueilli auprès des personnes concernées. L'obligation de rapporter la preuve du consentement ne peut pas être remplie par la seule présence d'une clause contractuelle engageant l'une des parties à recueillir un consentement valable pour le compte de l'autre partie. En effet, une telle clause ne permet pas à l'organisme de garantir, en toutes circonstances, l'existence d'un consentement valide (voir la délibération de la formation restreinte de la CNIL n° SAN-2023-009 du 15 juin 2023). Le contrat pourra, en revanche, être utilisé pour encadrer :

- les mécanismes mis en place pour permettre de démontrer le recueil d'un consentement valide ;
- la mise à disposition des éléments de preuve au profit de l'organisme qui souhaite se prévaloir du consentement ;
- le cas échéant, les conditions dans lesquelles ces éléments de preuve doivent être conservés, notamment afin de conserver leur valeur probante.

Exemple : le fournisseur d'un système d'IA générative d'image se rapproche d'un courtier en données pour constituer une base de données à des fins d'apprentissage comportant notamment des photographies.

Ils concluent pour cela un contrat qui garantit au fournisseur la licéité des données partagées, et encadre la fourniture d'indications cruciales pour la conformité de ses traitements (par exemple : preuves du contexte de la collecte des données pour apprécier son intérêt légitime, garanties s'agissant d'autres réglementations comme celle régissant la cession des droits de propriété intellectuelle, etc.).

Outre ces vérifications préalables, et quel que soit le mode de collecte utilisé, **les réutilisateurs doivent s'assurer de la conformité complète de leurs propres traitements.**

A noter que cette obligation vaut également lorsqu'ils réutilisent des bades de données dont la constitution et la diffusion ne relèvent pas du droit français ou européen. Pour plus d'informations sur le champ d'application territorial du RGPD, voir la fiche « Quel est le périmètre des fiches pratiques sur l'IA ».

En particulier, le réutilisateur doit veiller au respect des exigences vis-à-vis des personnes dont les données sont présentes dans la base ainsi obtenue : il doit les informer du traitement qu'il souhaite faire des données et leur permettre d'exercer leurs droits.

À noter : une fiche cas d'usage sur la réutilisation des données personnelles sera publiée ultérieurement. Elle permettra de compléter les éléments de mise en conformité introduits dans cette fiche notamment par l'étude de cas pratiques.

< Fiche précédente : Assurer que le traitement est licite - Définir une base légale

Sommaire

Fiche suivante : Réaliser une analyse d'impact si nécessaire >