

IA : Assurer que le traitement est licite - Définir une base légale

08 avril 2024

L'organisme qui souhaite constituer une base de données d'apprentissage contenant des données personnelles puis l'utiliser pour entraîner un algorithme doit s'assurer que le traitement est autorisé par la loi. La CNIL vous aide à déterminer vos obligations en fonction de votre responsabilité et des modalités de collecte ou de réutilisation des données.

Le responsable de traitement doit définir, dans tous les cas, une [base légale](#) et doit effectuer, en fonction du mode de collecte ou de réutilisation des données, certaines vérifications supplémentaires.

Il existe plusieurs moyens de constituer une base de données à des fins d'apprentissage, qui peuvent être utilisés de manière cumulative :

- les données sont collectées directement auprès des personnes ;
- les données sont collectées à partir de sources ouvertes sur Internet pour cette finalité ;
- les données ont initialement été collectées pour un autre objectif par le responsable du traitement lui-même (par exemple, dans le cadre de la fourniture d'un service à ses utilisateurs) ou par un autre responsable de traitement. **Cela implique de prendre des précautions complémentaires.**

Définir une base légale

Le principe

Comme tout traitement, la constitution et l'utilisation d'une base de données pour l'entraînement de modèle ou le développement de systèmes d'IA contenant des données personnelles ne pourra être mise en œuvre que si elle correspond à l'une des « bases légales » prévues par le RGPD.

Concrètement, la base légale est ce qui donne le droit à un organisme de traiter des données personnelles. Le choix de cette base légale est donc une première étape indispensable pour assurer la conformité du traitement. Selon celle qui sera retenue, les obligations de l'organisme et les droits des personnes pourront varier.

Les bases légales les plus pertinentes pour l'entraînement d'un algorithme sont détaillées ci-après.

En pratique

La détermination de la base légale doit être effectuée de manière adaptée à la situation et au type de traitement. Pour constituer une base de données pour l'entraînement d'un système d'IA, **les bases légales suivantes peuvent notamment être envisagées.**

La base légale du consentement

Pour être valide, [le consentement des personnes concernées](#) doit remplir quatre critères cumulatifs : **il doit être libre, spécifique, éclairé et univoque**. Le responsable du traitement doit être en mesure de démontrer la validité du recours à cette base légale en s'assurant que chacune de ces conditions, précisément définies par le RGPD, est remplie.

Exemple : un organisme souhaite filmer ou photographier des volontaires pour constituer une base de données d'images permettant d'entraîner un système à détecter certains gestes spécifiques. Il peut fonder le traitement sur la base de leur consentement.

Lorsqu'il constitue une base de données pour l'apprentissage d'un modèle d'IA, un organisme doit s'assurer **du consentement** recueilli.

Au-delà des obligations liées à la transparence, le consentement doit être accompagné d'un certain nombre d'informations communiquées à la personne avant qu'elle ne consente afin de lui permettre de prendre des décisions en toute connaissance de cause et de savoir comment retirer son consentement.

Le consentement doit porter sur une finalité spécifique ([voir la fiche n°2 sur la définition de la finalité](#)).

La liberté du consentement implique, en principe, de garantir la possibilité pour les personnes concernées de donner leur consentement de manière granulaire, lorsque les finalités poursuivies sont distinctes.

Exemple : le consentement de personnes à l'utilisation de leur image, collectée lors d'un événement d'une entreprise à des fins de communication, ne signifie pas qu'ils consentent à une réutilisation des données pour la constitution d'une base de données d'apprentissage ou d'amélioration d'un système d'IA. Dans ce cas, deux consentements distincts doivent être recueillis (par exemple au moyen de deux cases à cocher).

La liberté du consentement doit également faire l'objet d'une certaine vigilance en cas de déséquilibre de rapports de force entre la personne concernée et le responsable du traitement, en particulier si ce dernier est une autorité publique ou un employeur.

Exemple : pour développer un système d'IA, une entreprise souhaite utiliser les données de ses salariés. Leur consentement ne peut alors être valablement recueilli que dans des situations exceptionnelles, lorsqu'ils sont en mesure de refuser de donner leur consentement sans craindre ou encourir de conséquences négatives. En tant que responsable du traitement, l'entreprise devra veiller, en tout état de cause, à ce que les communications destinées à présenter le dispositif aux salariés ne soient ni incitatives,

ni contraignantes. Elle devra informer les volontaires de la possibilité de ne plus participer à la collecte de leurs données à tout moment, sans que cela porte à conséquence pour ces derniers.

Il n'apparaît pas possible de recueillir un consentement valide dans certains cas. C'est souvent le cas lorsque le responsable du traitement collecte des données accessibles en ligne ou réutilise une base de données ouverte, compte tenu notamment de l'absence de contact avec les personnes concernées et de la difficulté à les identifier. Dans ces cas de figure le responsable du traitement doit mobiliser une autre base légale plus adaptée.

Il peut également exister des difficultés liées au droit de retirer son consentement, par exemple du fait d'obstacles techniques à l'identification des personnes concernées. S'il n'est pas possible, pour le responsable du traitement, de garantir la possibilité d'exercer ce droit, il est recommandé de se fonder sur une base légale.

La base légale de l'intérêt légitime

[L'intérêt légitime](#) du responsable de traitement peut être retenu sous réserve du respect des conditions suivantes :

- **la légitimité de l'intérêt poursuivi par le responsable de traitement.** Par exemple l'intérêt pour un organisme de développer un modèle en vue de la commercialisation d'un système d'IA ou encore en vue de contribuer à l'amélioration de la connaissance scientifique, par exemple par la publication des outils développés (code, modèle, protocole expérimental, etc.) et des résultats de recherche.
- **la nécessité du traitement de données.** Par exemple, le traitement à des fins de constitution d'une base de données d'apprentissage contenant des images de personnes peut être considéré comme nécessaire aux intérêts d'un organisme qui souhaite développer un système de [détection de pose](#), lorsque des données anonymes ou synthétiques ne suffisent pas.
- **l'absence d'atteinte disproportionnée aux intérêts et droits des personnes concernées,** compte tenu de leurs attentes raisonnables à l'égard de ce traitement. La mise en balance des droits et intérêts en cause dépend des caractéristiques concrètes du traitement envisagé et notamment des garanties mises en œuvre pour assurer le meilleur équilibre possible entre ces intérêts et limiter les impacts du traitement sur les personnes concernées.

Le plus souvent, le fait de constituer une base de données pour l'entraînement d'un modèle dont l'usage est lui-même légal peut être regardé comme légitime. Une analyse au cas par cas est cependant nécessaire pour déterminer si l'utilisation de données personnelles à cette fin ne porte pas une atteinte disproportionnée à la vie privée des personnes concernées, et ce même lorsque les données ne sont pas nominatives. Pour assurer que son traitement est proportionné, le responsable de traitement peut notamment recourir à des mesures telles que la pseudonymisation des données, garantir l'absence de données sensibles, définir des critères de sélection permettant de limiter la collecte aux données pertinentes et nécessaires pour le traitement, etc.

Exemples :

Une entreprise souhaite développer un système d'IA capable de prédire le profil psychologique d'une personne à partir de données accessibles en ligne susceptibles de la concerner. Son intérêt commercial à développer un tel système sera vraisemblablement insuffisant au regard des intérêts, droits et libertés des personnes concernées : une autre base légale devra être recherchée ou le projet abandonné.

Un organisme constitue une base de données d'apprentissage en collectant les commentaires rendus publics et librement accessibles par des utilisateurs en ligne sur des forums, blogs et sites web. La finalité de ce traitement est de concevoir un système d'IA permettant d'évaluer et de prévoir l'appréciation d'œuvres d'art par le grand public. Dans ce cas, son intérêt à développer et éventuellement commercialiser un système d'IA peut être considéré comme légitime. La collecte de commentaires d'appréciation sur les œuvres peut être considérée comme nécessaire pour le développement du modèle, notamment compte tenu de la quantité de données requises pour l'apprentissage. Il convient de noter que la base légale de l'intérêt légitime donne le droit aux personnes concernées de s'opposer au traitement de leurs données (pour des motifs tenant à leur situation particulière).

Pour en savoir plus : fiche sur la mobilisation de l'intérêt légitime pour le développement des systèmes d'IA qui sera publiée prochainement.

La base légale de la mission d'intérêt public

La possibilité de se fonder sur la base légale de la « mission d'intérêt public » suppose :

- que la mission dans laquelle s'inscrit le traitement soit prévue par un texte normatif applicable au responsable du traitement ;
- que l'utilisation des données permette d'exercer spécifiquement cette mission de manière pertinente et appropriée.

Exemples :

Les chercheurs d'un laboratoire de recherche public sur la langue française souhaitent analyser l'évolution de l'utilisation de la langue en ligne. Ils constituent pour cela une base de données à partir des commentaires publiés librement en ligne sur différents réseaux sociaux (anonymisés à bref délai) afin d'entraîner un modèle qui détecte et analyse automatiquement l'occurrence de certaines expressions ou formes orthographiques.

Dans la mesure où le responsable de traitement est un laboratoire public, les chercheurs peuvent dans ce cas fonder le traitement de données sur la mission d'intérêt public. Cette base légale peut être mobilisée, de manière générale, pour les traitements de données effectués par des laboratoires de recherche publics ou privés investis d'une

mission d'intérêt public, dont les traitements de données sont nécessaires pour leur activité de recherche.

Le pôle d'expertise de la régulation numérique (PEReN) est autorisé à réutiliser, dans certaines conditions, des données publiquement accessibles de certaines plateformes afin de réaliser des expérimentations ayant notamment pour objet de concevoir des outils techniques destinés à la régulation des opérateurs de plateformes en ligne, conformément à l'article 36 de la [loi n° 2021-1382 du 25 octobre 2021](#) et au [décret n° 2022-603 du 21 avril 2022](#).

Pour plus d'information :

- Fiche cas d'usage n° 4 du [Guide sur la réutilisation des données publiquement accessible](#) (open data)
- [Quelle base légale pour un traitement de recherche ?](#)

La base légale du contrat

La base légale du contrat pourrait être mobilisée pour la constitution d'une base de données d'apprentissage d'un système d'IA à condition, d'une part, qu'un contrat valide soit conclu entre le responsable et la personne concernée et, d'autre part, que le traitement soit objectivement nécessaire à son exécution.

Les contrats conclus à cette fin doivent respecter les autres règles applicables, en matière de droit du travail ou de propriété intellectuelle par exemple.

Exemples :

Un éditeur de logiciel de traitement de texte propose un service de génération automatisée et personnalisée de courriers, auquel l'utilisateur souscrit contractuellement, et pour lequel il collecte les données des utilisateurs bénéficiaires de ce service.. Le traitement des données pour ce service de personnalisation peut être considéré, sous réserve des caractéristiques spécifiques du traitement de données, nécessaire à l'exécution du contrat.

À l'inverse, l'opérateur d'un réseau social en ligne inscrit dans ses conditions générales d'utilisation qu'il entend réutiliser les données de ses utilisateurs ([fournies par ces derniers, observées ou déduites par l'opérateur](#)) pour développer et améliorer de nouveaux produits, services et fonctionnalités utiles pour ses utilisateurs. Il ne peut pas fonder le traitement sur la base légale du contrat dès lors que ce traitement n'est pas objectivement indispensable pour leur offrir son service de réseau social en ligne (CJUE, 4 juillet 2023, Meta Platforms Inc. et a. c/Bundeskartellamt, [C-252/21](#)).

Données sensibles : un traitement interdit, sauf exceptions

[Les données sensibles](#) sont une catégorie particulière de données personnelles définies à l'article 9 du RGPD. Constituent par exemple des données sensibles des données qui révèlent la prétendue origine raciale ou ethnique des personnes concernées, ou encore des données biométriques aux fins d'identifier une personne physique de manière unique, comme un [gabarit facial](#) par exemple.

Le RGPD interdit le traitement de ces données, sauf exception, seulement dans les cas énumérés dans son article 9.2. Ces exceptions incluent notamment :

- les traitements pour lesquels la personne concernée a donné son consentement explicite (démarche active, explicite et de préférence écrite, qui doit être libre, spécifique, et informée) ;
- les traitements portant sur des données personnelles qui sont manifestement rendues publiques par la personne concernée ;
Dans ses lignes directrices sur le ciblage des utilisateurs des réseaux sociaux, le CEPD fournit une liste d'éléments à prendre en compte pour déterminer si les données sont manifestement rendues publiques : le paramétrage par défaut de la plateforme de réseaux sociaux, la nature de la plateforme, l'accessibilité de la page concernée, la visibilité de l'information sur leur caractère public, le point de savoir si la personne concernée a elle-même publié les données ou si elles l'ont été par un tiers ou déduites.
Il importe de vérifier si la personne concernée souhaitait, de manière explicite et par un acte positif clair, sur la base d'un paramétrage effectué en connaissance de cause, rendre accessibles au grand public ses données personnelles ou, au contraire, à un nombre plus ou moins limité de personnes sélectionnées (CJUE, 4 juillet 2023, [Meta Platforms](#), C-252/21).
- les traitements nécessaires pour des motifs d'intérêt public important, sur la base du droit de l'UE ou d'un État membre ;
- les traitements nécessaires à des fins de recherche scientifique sur la base du droit de l'Union européenne ou d'un État membre

Il convient de faire preuve d'une attention particulière **à la collecte de données sensibles lors de l'utilisation d'outils de moissonnage (*web scraping*)** qui impliquent le traitement de larges volumes de données. Le responsable du traitement est tenu de mettre en œuvre toutes les mesures permettant d'exclure automatiquement la collecte des données sensibles non pertinentes notamment notamment en appliquant des filtres permettant d'exclure la collecte de certaines catégories de données ou encore d'exclure certains sites comportant des données sensibles par nature.

Si, malgré les mesures prises, l'organismes traite de manière incidente et résiduelle des données sensibles qu'il n'avait pas cherché à collecter, cela n'est pas considéré comme illégal. C'est notamment ce qu'a pu considérer la cour de justice de l'Union européenne en rappelant que cette interdiction s'applique à l'exploitant d'un moteur de recherche « *dans le cadre de ses responsabilités, de ses compétences et de ses possibilités* » (CJUE, grande chambre, 24 septembre 2019, GC e.a, [C-136/17](#)). En revanche, si l'organisme vient à savoir qu'il traite des données sensibles, il est tenu de procéder, autant que possible, à leur suppression immédiate et automatisée de la base de données.

A noter :

- Une fiche sur la gestion des biais sera publiée ultérieurement. Elle permettra d'éclaircir la possibilité de traiter des données sensibles à des fins de détection et de correction de biais dans la base de données d'apprentissage.
- La CNIL mène actuellement des travaux sur la question de l'IA dans le domaine de la santé, qui feront l'objet d'une publication ultérieure.

La base de l'obligation légale

Si cette base légale peut sembler pertinente dans certains cas pour les traitements de données effectués en phase de déploiement, dans la mesure où l'utilisation d'un système d'IA peut parfois servir au responsable du traitement pour respecter une obligation légale (à condition de démontrer que celle-ci impose un traitement de données personnelles), elle est, en revanche, plus difficile à mobiliser pour fonder son développement.

En effet, pour mobiliser cette base légale, le responsable du traitement doit démontrer en quoi son traitement est nécessaire pour répondre à une obligation légale déterminée à laquelle il est soumis. Le texte sur lequel elle repose doit au moins définir la finalité du traitement et peut l'encadrer de manière plus précise (notamment à travers les types de données à traiter, la limitation des finalités ou d'autres conditions à respecter). Plus l'obligation légale est précise, plus il est facile de justifier en quoi elle impose un traitement de données personnelles.

Toutefois, les obligations n'étant généralement pas suffisamment précises pour prévoir le développement de systèmes d'IA, il conviendra le plus souvent de se fonder sur une autre base légale pour développer ce type de système.

Exemples :

Dans le secteur de l'assurance, les études actuarielles reposent sur des modélisations mathématiques, probabilistes et statistiques assimilables à des systèmes d'IA, dont l'objectif est de permettre d'identifier, qualifier et quantifier des risques (ainsi que les montants associés) liés aux contrats d'assurance.

Or, les obligations générales de solvabilité des organismes d'assurance n'étant pas suffisamment précises, il n'est pas possible de considérer que le développement de tels systèmes soit nécessaire à leur respect. L'intérêt légitime apparaît alors comme la base légale la plus pertinente.

[< Fiche précédente : Déterminer la qualification juridique des fournisseurs de systèmes d'IA](#)

[Sommaire](#)

[Fiche suivante : Assurer que le traitement est licite - En cas de réutilisation des données >](#)