CONSULTATION PUBLIQUE

FICHES PRATIQUES IA SUR L'ANNOTATION DES DONNEES ET LA SECURITE DU DEVELOPPEMENT DE SYSTEMES D'INTELLIGENCE ARTIFICIELLE

SYNTHESE DES CONTRIBUTIONS

Juin 2025

La CNIL a lancé, le 10 juin 2024, une consultation publique sur le développement de systèmes d'IA.

Les contributions ont nourri les travaux du deuxième lot de huit fiches pratiques, dont celles sur l'annotation et la sécurité du développement des systèmes d'intelligence artificielle pour leur publication définitive sur le site de la CNIL.

La synthèse en chiffres

Le 8 avril 2024, à la suite d'une consultation publique, **la CNIL a publié [un premier lot de fiches pratiques sur le développement des systèmes d'IA](#)** qui répondent en particulier aux interrogations portant sur l'application des principes de finalité, de minimisation et de durée de conservation pour le développement des systèmes d'IA et précisent certaines règles applicables à la recherche scientifique, à la réutilisation de bases de données ou à la réalisation d'analyse d'impact sur la protection des données (AIPD).

Dans la continuité de ces travaux, **la CNIL a lancé, le 10 juin 2024, [une nouvelle consultation publique sur un questionnaire concernant l'application du RGPD aux modèles d'IA et les 7 fiches suivantes](#)** :

- Base légale de l'intérêt légitime et développement de systèmes d'IA.
- Intérêt légitime : focus sur la diffusion des modèles en source ouverte (open source).
- Intérêt légitime : focus sur le moissonnage (*web moissonnage*).
- Informer les personnes concernées.
- Respecter et faciliter l'exercice des droits des personnes concernées.
- Annoter les données.
- Garantir la sécurité du développement d'un système d'IA.

À l'issue de la consultation publique, le 1^{er} octobre 2024, la CNIL a reçu 62 contributions représentatives de différents secteurs :

- des organismes à but lucratif :
 - 28 entreprises privées, dont des fournisseurs de systèmes d'IA, des utilisateurs de systèmes d'IA, ou encore d'autres entreprises de différents secteurs (IA, finance, santé, aéronautique, opérateurs de plateformes en ligne, publicité en ligne, audiovisuel, etc.) ;
 - 11 organisations représentantes des professionnels ;
 - 4 cabinets d'avocat ou de conseil.
- des organismes à but non-lucratif :
 - 3 associations représentatives de la société civile ;
 - 5 chercheurs ou groupement de chercheurs académiques et 1 think tank ;
- 6 particuliers ;
- 4 administrations publiques.

Cette synthèse se concentre sur les retours qui ont concerné les fiches « Annoter les données » et « Garantir la sécurité du développement d'un système d'IA ».

Ces contributions ont permis à la CNIL :

- de faire évoluer ses projets de fiches en apportant des clarifications supplémentaires et consolider ses analyses ;
- d'apporter des réponses, dans la synthèse ci-dessous, aux préoccupations les plus fréquemment partagées par les contributeurs.

Fiche n°11 – Annoter les données

1. Périmètre des recommandations

Sur l'objectif des recommandations concernant l'annotation

Des questionnements sur les cas visés par certaines des recommandations ont été relevés parmi les réponses. Il est à noter que les recommandations formulées dans cette fiche visent à couvrir une grande diversité de cas pour lesquels chacune des mesures proposées n'est pas toujours nécessaire. Conformément au principe de responsabilisation, il revient au responsable de traitement d'évaluer, pour son cas, les risques spécifiques pertinents et leur ampleur, et ainsi d'identifier les recommandations permettant d'en garantir le contrôle. Ainsi, le détail de l'information fournie aux personnes concernées, ou encore la rigueur du protocole d'annotation dépendront du contexte du traitement, en tenant compte par exemple de la nature des données traitées, des risques pour les personnes concernées ou encore de la finalité du traitement.

Il doit être précisé que bien que la phase d'annotation puisse présenter des risques qu'il convient de maîtriser, cette fiche ne vise pas à promouvoir davantage le recours à une approche non-supervisée. En effet, ce type d'entraînement de modèle n'est pas pertinent dans tous les cas, et il comporte également ses propres risques. Le choix du type d'apprentissage revient ainsi au responsable de traitement selon les spécificités techniques de son traitement.

Enfin, bien qu'un lien important existe entre la qualité de l'annotation et la performance du modèle, les recommandations de la CNIL n'ont pas vocation à conseiller les responsables de traitement dans l'atteinte d'une meilleure performance pour leur traitement. Pour cette raison, la question de l'impact de l'annotation sur la performance n'est évoquée que lorsque celle-ci peut avoir un impact sur les individus, tel que l'introduction d'un biais discriminatoire ou des effets physiques, psychologiques, juridiques, etc.

L'application de la fiche en pratique

Une réponse a indiqué que la mise en œuvre des recommandations de la fiche pouvait entraîner une charge importante dans la création ou l'amélioration d'un processus d'annotation, en ayant recours à un nouveau prestataire, ou encore pour la mise en place de nouveaux projets d'annotations par exemple. Tout d'abord, et comme rappelé ci-dessus, il revient au responsable de traitement d'évaluer le niveau de risque de chacun de ses traitements d'annotation et des modifications qui leur sont apportées afin de déterminer les mesures applicables. Ainsi, les mesures les plus lourdes pourront n'être mises en œuvre que lorsqu'elles sont adaptées au niveau de risque du traitement, à moins qu'il ne s'agisse d'une obligation légale (comme l'information, bien que son contenu puisse varier, notamment en application des bonnes pratiques recommandées dans la fiche). De plus, nombre des mesures recommandées dans la fiche peuvent représenter une charge lors de leur mise en œuvre, puis perdurer sans charge supplémentaire, à l'instar de la création d'un protocole d'annotation pouvant servir de modèle pour de nouvelles annotations, ou encore de la création d'une procédure pour l'exercice des droits.

Les annotations qualifiables de données à caractère personnel

Concernant la qualification des annotations, il a été soulevé que les annotations ne contenaient souvent pas de données à caractère personnel. Comme précisé dans la fiche, les annotations peuvent être qualifiées de données à caractère personnel et ainsi être dans le périmètre des recommandations de la CNIL dans deux situations :

- lorsqu'elles contiennent des données à caractère personnel, c'est-à-dire lorsque l'annotation contient des informations permettant de réidentifier un individu, comme par exemple son nom ou un identifiant ;
- lorsqu'elles portent sur des données à caractère personnel, puisqu'il est possible de relier cette annotation (qui comporte généralement des informations supplémentaires à celles contenues dans la donnée sans annotation) à la donnée.

Ainsi, bien qu'une annotation ne contienne pas toujours de données à caractère personnel, la possibilité de la relier à la donnée visée par l'annotation et contenant des données personnelles implique qu'il s'agit bien d'une information se rapportant à une personne physique identifiée ou identifiable. L'annotation pourra alors à son tour être qualifiée de donnée à caractère personnel.

Articulation avec le règlement européen sur l'IA

Certaines réponses ont suggéré que les recommandations sur l'annotation ne portent que sur les systèmes à haut risque désignés par le Règlement européen sur l'IA. Toutefois, l'évaluation des risques effectuée dans ce règlement n'est pas toujours la même que celle introduite par le RGPD. Il est ainsi possible de considérer des modèles d'IA dont l'entraînement peut être considéré comme un traitement à risque élevé pour le RGPD, et aboutissant à un niveau de risque minime pour le Règlement IA ou n'est pas dans son périmètre (comme par exemple un modèle entraîné sur des données sensibles concernant des personnes vulnérables, mais n'ayant pas vocation à être mis sur le marché). Ainsi, les périmètres des fiches et du règlement IA ont vocation à rester distincts. Ceci vaut notamment pour les dispositions prévues à l'article 10.2, requérant que les jeux de données d'entraînement, de validation et de test soient soumis à des pratiques en matière de gouvernance et de gestion des données appropriées à la destination du système d'IA à haut risque, notamment en ce qui concerne l'annotation.

2. Le respect des principes d'exactitude et de minimisation

En ce qui concerne l'exactitude des données et de leurs labels, il a été soulevé que celle-ci pouvait représenter une difficulté en raison de l'absence de certitude sur l'exactitude des données lors de leur collecte. Comme indiqué à l'article 4 de la loi Informatique et Libertés, le principe d'exactitude exige que toutes les mesures raisonnables soient prises pour effacer ou rectifier les données inexactes eu égard aux finalités du traitement. Ces mesures étant prises lors de la collecte, l'exactitude des annotations ne peut chercher à corriger une inexactitude des données, à moins que celle-ci ne soit identifiée lors du processus d'annotation. Ainsi, l'exactitude des annotations dépend de l'exactitude des données et exige, elle aussi, la mise en œuvre de toutes les mesures raisonnables eu égard à la finalité.

Il est à noter qu'une annotation pouvant porter atteinte à la réputation de la personne (comme une annotation dégradant son image) pourra toutefois être exacte, et justifiée au regard de la finalité du traitement. Cela peut être le cas par exemple dans la lutte contre la criminalité.

3. Le protocole d'annotation

Certaines réponses ont souligné la charge induite par la création d'un protocole d'annotation tel que proposé dans la fiche. Il doit être rappelé qu'il s'agit d'une pratique courante, d'ores et déjà mise en œuvre par les développeurs de systèmes d'IA pour garantir la qualité des annotations, et à laquelle viennent s'ajouter les recommandations de la CNIL visant à protéger les droits des personnes concernées en tenant compte du niveau de risque du traitement. Ainsi, bien que ces procédures représentent une charge, celle-ci n'est pas additionnelle aux tâches des développeurs de modèles, et comporte un double intérêt, tant pour les personnes concernées que pour le responsable de traitement.

Le choix des labels

Si celui-ci doit tenir compte des enjeux d'exactitude et de minimisation afin d'éviter notamment une atteinte à la réputation des personnes, l'objectif premier reste l'atteinte de la finalité du traitement lorsqu'il est licite et proportionné. Ainsi, les labels sélectionnés doivent en premier lieu permettre d'entraîner le modèle d'IA, à condition qu'ils ne remettent pas en cause la licéité du traitement. Cet objectif sera par ailleurs souvent concomitant avec la protection des droits, en évitant par exemple les annotations ambiguës, ou inexactes.

Il a été soulevé que l'objectivité lors de l'annotation pouvait être difficile à atteindre dans certains cas, notamment lorsque l'annotation est réalisée par une personne sur la base de critères objectifs mais laissant une marge d'appréciation. Dans ces cas, l'objectivité doit en effet être tenue comme un but, et non une contrainte stricte puisqu'il ne serait pas possible de la respecter en pratique.

La vérification continue

Certaines contributions ont mentionné des outils permettant de vérifier la qualité de l'annotation, en identifiant par exemple la présence de préjugés sexistes dans des corpus de texte. Ces outils peuvent en effet apporter une aide précieuse pour la mise en œuvre des recommandations. Leur qualité doit toutefois être appréciée au cas par cas, par exemple en sélectionnant les outils à l'état de l'art et en considérant leurs limitations inhérentes, notamment pour l'évaluation de critères éthiques, pour laquelle une approche technique ne sera pas suffisante. En particulier, les métriques visant à évaluer l'équité des annotations reposent nécessairement sur les choix de leurs concepteurs et possèdent des limitations, telles que l'évaluation de biais intersectionnels.

L'implication d'un référent ou d'un comité éthique

Il est souligné dans certaines réponses que la supervision de l'annotation par un référent ou un comité éthique pourrait ajouter une charge organisationnelle importante. Le rôle de ce référent ou comité peut être crucial pour le respect de certains critères éthiques, et en particulier pour assurer que l'annotation portant sur les individus ne porte pas atteinte à leur droits, et évite l'apparition de biais discriminatoires. Cette supervision spécifique à l'IA est justifiée par le risque d'automatisation d'atteinte aux droits des personnes dans le domaine. L'implication d'un référent ou comité reste toutefois à apprécier selon la situation et ne sera pas toujours pertinente.

4. L'information

Plusieurs réponses ont soulevé que les bonnes pratiques relatives à l'information sur l'annotation pouvaient être trop exigeantes. Comme indiqué dans la fiche, ces recommandations sont données à titre de bonne pratique et ne seront pas applicables dans tous les cas. En ce qui concerne spécifiquement les critères de responsabilité sociale des entreprises d'annotation, cette information pouvant être utile aux personnes concernées afin d'évaluer les risques et garanties liés au traitement (considérant 39 du RGPD) et ainsi de décider de l'exercice de leurs droits, il reste recommandé d'en faire mention dans l'information qui leur est fournie. Cette information n'étant pas pertinente dans tous les cas, il peut être plus utile de privilégier la transmission d'autres informations afin de privilégier l'intelligibilité pour les personnes concernées.

5. Autres remarques

L'impact de l'annotation sur l'anonymisation des données

Une réponse a pointé que l'annotation pouvait augmenter le risque de réidentification des individus en reliant les données d'une même personne entre elles (via un identifiant, ou un horodatage permettant de rapprocher des données entre elles). Ce risque accru est en effet pertinent lorsque l'annotation porte sur des données qu'on cherche à anonymiser.

La responsabilité du traitement d'annotation

Pour certains participants, la répartition des responsabilités de traitement d'annotation demande à être clarifiée.

Dans de nombreux cas, l'organisme réalisant l'annotation sera un prestataire au rôle de sous-traitant. Cela sera notamment le cas lorsqu'il ne décide pas des données devant être annotées, des labels ou encore de la destination des données annotées. Dans ce cas, le responsable de traitement devra mettre en œuvre les recommandations de la CNIL en lien avec le sous-traitant.

Dans d'autres situations, l'organisme réalisant l'annotation pourra être responsable de traitement. Cela sera le cas lorsqu'il décide des moyens et des finalités du traitement, comme par exemple lorsqu'un organisme collecte les données, les annoté et les partage à un autre organisme. Alors il pourra ne pas être responsable du traitement réalisé par le destinataire des données. Dans ce cas, seules certaines recommandations s'appliqueront à lui alors qu'il ne sera pas en mesure d'en réaliser d'autres (notamment portant sur l'évaluation du risque de régurgitation et d'inférence comme décrit ci-après).

La responsabilité de l'évaluation du risque de régurgitation et d'inférence des données sensibles

Il a été soulevé que l'évaluation du risque de régurgitation et d'inférence sur les données partagées par le responsable du traitement d'annotation à partir d'un modèle entraîné sur ces données par un autre organisme n'était pas de la responsabilité de l'organisme réalisant l'annotation. En effet, celui-ci ne possède pas les informations nécessaires pour analyser ce risque (tel que le type de modèle entraîné, ou la stratégie d'apprentissage). Toutefois, il possède une connaissance sur le jeu de données, et notamment sur sa distribution statistique (comme la présence d'*outliers* et de données dupliquées). Ainsi, il peut être à même de fournir des informations utiles au développeur du modèle pour estimer le risque de régurgitation. Bien qu'il ne s'agisse pas d'une obligation, cette transmission d'information est recommandée à titre de bonne pratique.

1. Périmètre et remarques générales

L'application de la fiche en pratique

Tout d'abord, plusieurs réponses ont indiqué que les mesures recommandées pouvaient être trop exigeantes, notamment pour certaines structures aux moyens limités. Il est rappelé à cet égard que l'article 32 du RGPD exige la mise en œuvre de mesures techniques et organisationnelles appropriées afin de garantir un niveau de sécurité adapté au risque, compte tenu 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, les mesures de sécurité recommandées dans cette fiche ne seront pas obligatoires dans le cas général, mais devront être sélectionnées selon les spécificités du contexte du traitement. Dans le cas du développement d systèmes d'IA, il pourra par exemple être tenu compte de la nature des données traitées, de la vulnérabilité des personnes concernées, ou encore de la destination du modèle entraîné (i.e. des modalités de son partage, ou des finalités prévues pour son utilisation le cas échéant). De même, d'autres mesures que celles citées pourraient être pertinentes afin de réduire le niveau de risque spécifique à un traitement. Le choix des mesures pertinentes et leur priorisation revient ainsi au responsable de traitement et au sous-traitant selon le principe de responsabilisation.

Il doit être souligné que de la connaissance de la CNIL, le thème de la sécurité de l'IA est un sujet évolutif et encore peu maîtrisé par les acteurs du domaine. Les professionnels du secteur semblent encore peu formés aux méthodologies de gestion des risques habituelles, alors que les risques, méthodes d'évaluation et mesures de sécurité spécifiques à l'IA sont encore au stade de la recherche. L'attention portée à la sécurité devrait ainsi être particulièrement importante dans ce domaine, surtout dans un contexte innovant où les avancées technologiques sont privilégiées au détriment des impacts potentiels pour les personnes.

Le périmètre de la fiche

L'ensemble des systèmes d'IA est visé par cette fiche. Il doit toutefois être mentionné que les risques pertinents dépendront parfois du type de système considéré, puisqu'il aura des conséquences sur la nature des données traitées, leur volume, leur durée de conservation etc. Les recommandations n'ont en revanche pas vocation à encadrer davantage certains types de systèmes.

Il a également été mentionné que la fiche ne faisait pas de cas particulier de l'ajustement des modèles, alors que ce cas concernera un plus grand nombre d'acteurs que le développement dans sa totalité. L'ensemble des recommandations de la fiche porte bien sur le cas de l'ajustement également, ce qui sera précisé dans la fiche.

La sécurité du développement et la sécurité dès la conception

Certaines réponses ont pointé que certaines mesures de sécurité prévues dans la fiche visaient davantage à sécuriser le déploiement que le traitement lié au développement. En effet, en application de l'article 25 du RGPD, il est recommandé de mettre en œuvre les mesures de protection des données dès la conception du traitement. Ainsi, certaines mesures sur lesquelles seul le concepteur du modèle a la main seront recommandées lors du développement (la responsabilité de leur mise en œuvre dépendra toutefois selon les cas). C'est notamment le cas des recommandations visant à faciliter le contrôle humain sur le système puisque prévoir une fonctionnalité logicielle permettant la mise en pause du système sera généralement à la main de son fournisseur. La décision d'utiliser cette fonctionnalité reviendra toutefois à son utilisateur.

En tout état de cause, lorsque les mesures portent spécifiquement sur la sécurisation du déploiement, elles seront abordées dans des recommandations ultérieures.

L'articulation avec le règlement IA

Certaines réponses ont relevé que les mesures de sécurité préconisées étaient différentes de celles prévues par le Règlement européen sur l'IA. Toutefois, l'évaluation des risques effectuée dans ce règlement n'est pas toujours la même que celle introduite par le RGPD. Il est ainsi possible de considérer des modèles d'IA dont l'entraînement peut être considéré comme un traitement à risque élevé pour le RGPD, et aboutissant à un niveau de risque minime pour le Règlement IA ou n'est pas dans son périmètre (comme par exemple un modèle entraîné sur des données sensibles concernant des personnes vulnérables, mais n'ayant pas vocation à être mis sur le marché). Ainsi, les périmètres des fiches et du règlement IA ont vocation à rester distincts. Ceci vaut notamment pour les dispositions relatives à la sécurité des systèmes, puisque les risques visés par le règlement

IA ne sont pas toujours les mêmes que ceux liés à la protection des données personnelles, en particulier lors du développement des systèmes.

Au contraire, certaines mesures permettent parfois de répondre aux exigences de sécurité des deux textes (notamment à l'article 15.5 du Règlement IA qui cible les attaques visant la confidentialité). Ces mesures pourront alors être préférées afin de réduire la charge des obligations. Cela peut notamment être le cas en ce qui concerne l'analyse de certains risques (comme prévu à l'article 27 du règlement IA par exemple en ce qui concerne l'analyse d'impact sur les droits fondamentaux) ou la documentation du système et de sa sécurité.

2. Les mesures de sécurité à envisager pour le développement d'un système d'IA

Les réponses apportées ont suggéré plusieurs ajouts pertinents tels que :

- la réalisation d'une nomenclature logicielle (*software bill of material* en anglais) ;
- la création d'un plan d'intervention et d'un plan de retour en arrière en cas d'incident ;
- le recours à du matériel sécurisé, comme les environnements d'exécution de confiance (*Trusted execution environments*, ou TEE) ;
- l'utilisation des technologies améliorant la confidentialité (ou *privacy enhancing technologies*, ou PETs) ;
- le respect de la durée de conservation des données afin de limiter la surface d'attaque temporelle.

En revanche, d'autres suggestions n'ont pas été suivies :

- l'implication du référent ou comité éthique dans la sécurité du système. Cette implication peut être utile lorsque le comité possède des compétences dans la sécurité des systèmes, mais cette compétence ne peut être attendue dans tous les cas ;
- le recours au chiffrement homomorphe. Cette technique étant impraticable dans de nombreux cas et encore au stade de la recherche, elle n'est pour le moment pas recommandée dans le cas général.

Des questionnements ont porté sur certaines recommandations spécifiques :

- une réponse a soulevé que le cloisonnement des données sensibles pouvait permettre de les trouver plus facilement, notamment pour les données provenant d'une base contenant aussi des données non sensibles. Ce risque doit en effet être pris en compte et mis en balance avec les protections apportées par le cloisonnement ;
- il a été souligné que la journalisation des jeux de données était une mesure exigeante d'un point de vue technique. Cette mesure sera parfois inadaptée à certains cas et devrait être mise en œuvre lorsque les modifications apportées aux données sont vraisemblables et peuvent avoir des conséquences graves (comme dans le cas d'un empoisonnement des données redouté).
- le lien avec les standards sur la sécurité a fait l'objet d'une interrogation. Les standards sur la sécurité généralistes comme la norme ISO 27001 est pleinement applicable. Par ailleurs, d'autres normes spécifiques à l'IA en lien avec le règlement IA ont fait l'objet d'une demande par la Commission Européenne (notamment sur le système de gestion de risques) et seront des documents de référence utiles lors de leur adoption.

3. Autres

Un grand nombre de références, dont certaines sont citées dans la fiche, recensent les risques et mesures de sécurité applicables à l'IA. Certaines réponses ont fait référence à ces documents. Bien qu'il ne soit pas possible de tous les lister ou d'y faire référence précisément, la mise en œuvre des recommandations de la CNIL pourra utilement s'appuyer sur ces documents lorsque leur qualité est vérifiée.