Note de cadrage sur l'usage de l'IA Générative au sein d'Inria

Référence Gédéi : 17929

Date : 12 mai 2025

Rédacteur : Jean-Frédéric Gerbeau

Contributeurs : CORUM[1]: Sylvain Petitjean ; DAFP : Catherine Gallet-Rybak, Elodie Limou, Alexandre Sicard ; DAJ : Eric Jaouen ; DGDI : David Rey ; DGDS : Karteek Alahari, Christophe Biernacki, Jean-Frédéric Gerbeau ; DPD : Anne Combe, Emilie Masson ; DSI : Florian Dufour, Thierry Murgue ; FSD : Didier Benza ; RSSI : Dominique Launay.

Objet : Consignes en matière d'usage d'outils d'IA générative

Diffusion : aucune restriction

1. Recommandations générales

Les plateformes d'IA, utilisant en particulier des techniques d'IA générative, par leurs performances remarquables dans divers usages, sont susceptibles d'assister efficacement les personnels, dans les fonctions scientifiques comme dans les fonctions d'appui. Les modèles d'IA font par ailleurs partie des activités de recherche de nombreuses équipes-projets Inria.

Il est donc important de rappeler que leur utilisation, dans un cadre responsable et légal, doit se faire en respectant un certain nombre de règles et de précautions. De nombreux établissements, en France et à l'étranger, ont produit des guides (voir Annexe 1). Nous résumons ici les points qui nous semblent essentiels.

Recommandations générales, pour les fonctions d'appui et les fonctions scientifiques

- Il est important de s'abstenir d'utiliser des outils d'IA générative, en particulier non souverains, pour des activités sensibles pouvant impacter d'autres personnes ou organisations. Par exemple : fuite d'idées dans un contexte d'évaluation par les pairs ; données utilisées dans le cadre d'un partenariat industriel.
- Le risque de perte de propriété intellectuelle est réel en cas d'envoi à l'IA générative d'une nouvelle méthode ou d'une nouvelle invention qui n'est pas encore protégée. Par ailleurs, au-delà de la propriété intellectuelle, il faut être conscient qu'en l'utilisant, on fournit à la plateforme d'IA des informations sur ses sujets d'activité, ce qui peut constituer en soi une matière précieuse.
- Il convient d'opter dans la mesure du possible pour des modèles suffisamment ouverts. Le degré d'ouverture pouvant se mesurer selon trois grands critères : disponibilité (du code source, des données d'entraînement, des poids du modèle, ...), documentation (du code, de l'architecture, ...) et accès (code téléchargeable, API ouverte, ...).
- Les systèmes d'IA générative ne bénéficient pas — en l'état actuel du droit — de la personnalité morale et ne peuvent donc pas avoir la qualité d'auteur. L'utilisation de leurs productions se fera donc toujours sous la responsabilité et la supervision de l'utilisateur qui

[1] Une liste d'abréviations est donnée en Annexe 5.

devra posséder les connaissances et compétences nécessaires pour en évaluer la qualité et la validité, les soumettre à une analyse critique et en identifier les biais. La réponse d'une IA générative ne saurait être citée comme source primaire permettant de soutenir une affirmation.

- Il est recommandé de toujours déclarer et documenter l'utilisation d'un outil d'IA générative, et en détailler les modalités de manière transparente en mentionnant *a minima* le modèle, la version, les stratégies d'interaction (prompts), la date, la justification, etc.
- Dans le cas où l'IA générative est utilisée en appui de l'aide à la décision, il faut traiter les réponses fournies comme un élément parmi d'autres, et en aucun cas comme l'unique élément sur lequel s'appuie une décision.
- Il faut être conscient qu'une utilisation massive et irréfléchie des technologies d'IA générative peut conduire à aggraver certains impacts sociétaux négatifs du numérique. Par exemple : paresse cognitive ; perte de compétences ; aggravation de l'empreinte environnementale du numérique.

Recommandations plus spécifiques dans un cadre scientifique[2]

- Les politiques en matière d'usage permis des systèmes d'IA générative dans différents contextes (publication dans une revue ou une conférence, évaluation par les pairs, proposition soumise à financement, ...) sont encore assez variables. Il convient donc de s'enquérir systématiquement des dispositions mises en place par les acteurs concernés avant toute utilisation. Certains éditeurs interdisent par exemple le recours à l'IA pour la génération d'images.
- L'application des règles de précautions s'adaptera à la criticité de l'application visée. Par exemple :
    - La génération de jeux de données synthétiques – permettant de s'affranchir des règles relatives à la protection de la vie privée – et le recours à des participants virtuels dans le cadre d'expérimentations sont deux applications de l'IA générative en recherche qui doivent être considérées comme critiques.
    - Selon son objet, la génération de code informatique peut également être critique (voir section 2). Par ailleurs, les avertissements généraux s'appliquent à l'usage de l'IA générative en programmation, en particulier le risque d'inexactitude ou de plagiat déguisé.
- Il faut être conscient qu'une utilisation massive et irréfléchie des technologies d'IA générative peut aggraver certaines tendances négatives actuelles du fonctionnement de la recherche, par exemple :
    - L'augmentation de la production scientifique globale conduisant à un engorgement des procédures de relecture et d'acceptation et à un allongement des délais de publication ;
    - La publication massive d'articles scientifiques à la validité douteuse basés sur des données fabriquées, falsifiées ou plagiées (phénomène des « usines à papier » ou « *paper mills* ») conduisant à une pollution de l'espace informationnel, à une dégradation du corpus de connaissances sur lequel s'appuient les futures recherches, et *in fine* à un affaiblissement du lien de confiance entre la science et la société.
- De manière générale, il est important de garder à l'esprit que les systèmes d'IA générative vont, par construction, à l'encontre des grands principes d'honnêteté, de transparence, de véracité et d'indépendance fondant la recherche responsable et l'intégrité scientifique :

[2] Des recommandations plus détaillées sont présentées dans une note de la Commission d'Évaluation en cours de préparation.

données d'entraînement peu documentées et à la légalité discutable ; réponses probabilistes, reproduisant les biais du corpus d'entraînement et gommant la paternité ; affabulations ; versionnage opaque et retrait de modèles, ce qui nuit à la reproductibilité ; etc.

RGPD et AI Act

Le Règlement général de protection des données (RGPD, 2018) et le Règlement sur l'intelligence artificielle (AI Act, 2024) prévoient un principe de transparence et une exigence de documentation pour les systèmes d'IA. L'AI Act fixe des exigences spécifiques dont le respect peut contribuer largement au respect des exigences du RGPD.

Le RGPD s'applique seul pour un traitement de données personnelles servant à développer ou utiliser un système d'IA n'étant pas soumis à l'AI Act. Par exemple : un système exclusivement développé à des fins de recherche scientifique.

Le RGPD et l'AI Act s'appliquent lorsqu'un système d'IA à haut risque nécessite des données personnelles pour son développement ou dans son déploiement. Par exemple : un système d'IA utilisé pour le tri automatique de CV.

On renvoie aux Annexes 3 et 4 pour plus de détails.

2. Consignes en matière de SSI et de PPST

Que ce soit pour un usage scientifique ou général, l'utilisation d'une plateforme d'IA au sein d'Inria doit avoir fait l'objet d'une décision d'autorisation préalable par l'établissement. Elle doit être cadrée contractuellement et homologuée en fixant un niveau de sensibilité maximum autorisé pour les données transmises[3].

L'utilisateur doit évaluer la sensibilité de la session qu'il s'apprête à établir avec une plateforme d'IA et la sensibilité des données transmises ne doit pas dépasser le niveau maximum autorisé pour cette plateforme. Il ne faut par exemple pas partager des données d'article avant publication, ou des informations ou données couvertes par un accord de confidentialité, ou encore des éléments d'un document clairement estampillé confidentiel[4].

L'évaluation de la sensibilité peut se faire en répondant à la question suivante : « que se passerait-il si la totalité de la session que je m'apprête à avoir avec la plateforme d'IA était rendue publique et associée à mon identité » ? Si la réponse à cette question est :

- **Aucun impact**, la session est **publique** ;
- **Impact limité**, moi, l'institut ou d'autres personnes éprouverions une gêne momentanée, mais sans plus, la sensibilité de la session est « **diffusion limitée** » ;
- **Impact important**, moi, l'institut ou d'autres personnes serions durablement affectés par une telle diffusion, la sensibilité de la session est « **confidentielle** » ;
- **Impact très important**, moi, l'institut ou d'autres personnes, nous serions définitivement affectés par un tel incident, la sensibilité de la session est « **diffusion restreinte** ».

Par exemple, la diffusion à une plateforme non autorisée d'éléments soumis à un accord de confidentialité aurait un impact important sur la personne qui réaliserait cette diffusion et si elle venait à être découverte (sanctions disciplinaires).

---

[3] Circulaire interministérielle N° 3415/SGDSN/AIST/PST du 7 novembre 2012, section 4: Obligations du responsable d'une unité de recherche ou de production « protégée ».

[4] Guide ANSSI PA 102 - Recommandations de sécurité pour un système d'IA générative, R34 : « Proscrire l'utilisation d'outils d'IA générative sur Internet pour un usage professionnel impliquant des données sensibles ». Voir également la règle 9 de la doctrine « Cloud au centre » de l'État.

Dans le cas d'implémentations internes à l'institut, seules des données d'un niveau de sensibilité maximum « diffusion limitée » peuvent servir à l'entraînement de la plateforme d'IA. Si les requêtes peuvent servir à l'entraînement de la plateforme, l'utilisateur doit pouvoir refuser cet entraînement, notamment dans le cas de données d'une sensibilité supérieure à « diffusion limitée ».

L'Annexe 2 « PPST » fournit des indications de classification en fonction des usages et de la localisation de la plateforme utilisée.

L'utilisation de l'API d'une plateforme d'IA et la mise en place de dépendances automatisées entre cette plateforme et des éléments du système d'information d'Inria, scientifique ou non, doit obligatoirement faire l'objet d'une homologation de sécurité spécifique.

Le contrat avec l'opérateur de la plateforme d'IA, ou ses conditions générales d'utilisation, doivent préciser au minimum les conditions suivantes :

- Interdiction de conserver les données transmises par un utilisateur au-delà de la session de cet utilisateur.
- Interdiction d'utiliser les données transmises afin d'améliorer/entraîner les modèles de la plateforme[5].

Concernant la génération de code informatique[6], il est proscrit :

- D'exécuter du code généré par une plateforme d'IA sans vérification préalable par un être humain.
- De solliciter la génération de code concernant des applications sensibles, en particulier :
    - Des modules de gestion de droit d'administration et de droits d'accès ;
    - Des modules cryptographiques ;
    - Des modules de traitement de données sensibles.

3. Consignes en matière d'achat

À titre provisoire, dans l'attente de la mise en place de solutions d'accès à des plateformes d'IA générative par la DSI, l'achat de services extérieurs est toléré quand leur usage est nécessaire à l'accomplissement d'un travail de recherche.

Les commandes de licences d'utilisation d'une plateforme d'IA, ou les demandes de remboursement, doivent être soumis pour avis, sous couvert de la voie hiérarchique, au FSD, au DGDI et au DGDS, et feront l'objet d'un accord du PDG. Les services sollicités pour de telles commandes ou remboursement doivent refuser de les traiter sans cet accord.

Le recours à un marché public (DINUM, UGAP, ...) ne dispense pas de cet accord, quels que soient les supports de financement (subvention ou ressources contractuelles).

Cet accord est également indispensable pour le règlement par l'agent sous forme de menues dépenses, dont on rappelle qu'elles ne peuvent se faire que dans un cadre strict d'application (dépense modérée, non régulière, etc.).

Le formulaire de demande d'accord est sur Numin, page Achats : Dépenses & Remboursements. Il doit être adressé à <assistantes_dgds@inria.fr>

---

[5] Recommandation R21 du guide PA 102 de l'ANSSI : « Proscrire le réentraînement du modèle d'IA en production ».
[6] cf. chapitre 5 PSSI IA, CLUSIF et recommandations ANSSI.

4. Cellule opérationnelle LLM@Inria

Missions

La DSI et le Service de Développement Technologique (SDT) de la DGDI ont mis en place une cellule opérationnelle en IA dont les missions sont :

- Disposer d'un cadre technique permettant à des utilisateurs internes d'expérimenter des modèles d'IA ;
- Organiser la remontée et l'évaluation des tests effectués, sous une forme équivalente à celle d'un carnet de laboratoire, permettant de garder une trace de ceux-ci ;
- Éclairer régulièrement la direction d'Inria sur les tests effectués, et lui proposer des cas d'usage internes, pour préparer des décisions de déploiement ou d'évolution de solutions basées sur l'IA générative, à l'usage d'un métier par exemple ;
- Établir des échanges réguliers avec des partenaires d'Inria, en particulier les universités, dans un objectif d'enrichissement mutuel. À date, une attention particulière sera accordée aux activités de l'Université Grenoble Alpes et de l'Université de Rennes.

La cellule se compose d'agents de la DSI et de la ligne métier « développement technologique ». Des utilisateurs expérimentateurs pourront également en faire partie.

Centralisation

Pour éviter une dispersion des efforts et des compétences, la cellule LLM a vocation à centraliser l'ensemble des expérimentations sur des modèles d'IA génératives menées dans l'institut. Toute démarche visant à déployer un modèle d'IA doit donc faire l'objet d'un signalement par mail à la cellule <cellule-llm@inria.fr>. Les équipes-projets menant des expérimentations à des fins de recherche pourront s'appuyer sur les travaux de la cellule et y contribuer.

La cellule LLM met en place un processus de remontée et d'évaluation des tests effectués, sur une forme similaire à celle d'un cahier de laboratoire. Tout expérimentateur faisant appel à l'infrastructure de test doit s'insérer dans ce processus et enrichir la base de tests.

Parmi les bénéfices attendus de cette centralisation, on peut citer :

- La constitution d'une base de connaissances permettant de suivre l'état de l'art et l'évolution des usages ;
- La montée en compétences et la formation d'équipes grâce à ces retours ;
- L'amélioration continue du fonctionnement de la cellule et de l'infrastructure mise en place, pour le bénéfice de l'ensemble de l'institut.

Infrastructures

La cellule LLM s'appuie sur l'équipe produit ABACA[7] pour avoir accès à une infrastructure permettant de tester, à l'échelle d'Inria, des solutions d'IA générative :

- Avant tout sur l'infrastructure ABACA existante (matériel, administration et orchestrateur) ;
- À la marge et temporairement, sur des matériels dédiés, notamment pour de l'expérimentation spécifique de ces matériels, dont l'hébergement est toutefois assuré par la DSI de la même façon qu'ABACA.

L'utilisation d'ABACA à ces fins d'expérimentation sera limitée à quelques pourcents de la capacité de calcul d'ABACA.

[7] ABACA est l'infrastructure nationale de moyens de calcul Inria.

Annexe 1 – Bibliographie

Réglementation et recommandations sur l'IA Générative

- Union Européenne :
  - (Prévue pour mai 2025) - AI Office - Code de pratique général sur l'IA
  - 12/2024 – EDPB - Avis de l'EDPB sur les modèles d'IA
  - 11/2024 – AI Office - Modèles d'IA à finalité générale dans la législation sur l'IA
  - 08/2024 – Commission européenne - Question réponses sur l'IA
  - 06/2024 - Parlement européen - AI Act
- France :
  - 02/2025 – CNIL - IA et RGPD : Nouvelles recommandations pour accompagner une innovation responsable
  - 07/2024 – CNIL - Questions réponses sur l'IA Act
  - 04/2024 – CNIL - Recommandations pour développement de systèmes d'IA
  - 04/2024 – ANSSI PA 102 - Recommandations de sécurité pour un système d'IA générative

IA Générative en recherche

- International :
  - 01/2025 - University of Edinburgh - The Only Way is Ethics: A Guide to Ethical Research with Large Language Models
  - 10/2024 - University of Edinburgh - Whitepaper on Ethical Research into Large Language Models
  - 09/2024 – EASE - The European Association of Science Editors - Recommendations on the Use of AI in Scholarly Communication
  - 09/2024 - Université du Quebec Montréal - Guide pour une utilisation responsable de l'IA générative dans la recherche universitaire
  - 04/2024 - Radboud University - Generative AI and research integrity
  - 04/2024 - Macquarie University (Sydney) and Queensland University (Brisbane) - A University Framework for the Responsible use of Generative AI in Research
  - 03/2024 - European Commission - Living guidelines on the responsible use of Generative AI in research
  - 03/2024 – ENRIO - European Network of Research Integrity Offices - Artificial intelligence transforms research but the integrity norms do not
  - 03/2024 - European Commission - Successful and timely uptake of Artificial Intelligence in science in the EU
  - 12/2023 - Cornell University - Generative AI in Academic Research
  - 12/2023 - European Commission - Harnessing the power of AI to accelerate discovery and foster innovation
  - 09/2023 - Unesco - Guidance for generative AI in education and research.
  - 09/2023 - Cancer Research UK - Guidance for researchers on the use of generative AI
- France :
  - 10/2024 – Université d'Orléans - Charte sur l'utilisation de l'Intelligence Artificielle au sein de l'Université d'Orléans
  - 10/2024 - Cirad - Usages des outils d'Intelligence Artificielle générative dans le domaine de la recherche
  - 09/2024 - Inrae - Recommandations pour l'usage des IA génératives comme assistant personnel au sein d'INRAE
  - 03/2024 - Diffusion du RAG de l'Université de Rennes (RAGaRenn)

Annexe 2 – PPST – Classification en fonction des cas d'usage

Le tableau liste les niveaux de classification en fonction des usages et de la localisation de la plateforme d'IA : Public, Diffusion Limitée (DL), Confidentiel (CO), Diffusion Restreinte (DR).

Cas d'usage	***Localisation plateforme d'IA***	***Classification maximum autorisée***
1. Rédaction – *Aide à la rédaction* – *Correction grammaticale* – *Traduction*	*Hors Europe*	*Public, DL*
	Europe	*CO*
	Interne	*DR*
2. Analyse de texte – *Synthèse de texte*	*Hors Europe*	*Public, DL*
	Europe	*CO*
	Interne	*DR*
3. Développement – *Génération de code* – *Explication de bugs* – *Revue de code*	*Hors Europe*	*Public, DL*
	Europe	*CO*
	Interne	*DR*
4. Recherche d'informations – *Recherche d'informations* – *Documentation sur des sujets nouveaux*	*Hors Europe*	*Public, DL*
	Europe	*CO*
	Interne	*DR*
6. Autres – *Générer de nouvelles idées* – *Génération d'images*	*Hors Europe*	*Public, DL*
	Europe	*CO*
	Interne	*DR*

Annexe 3 – RGPD et systèmes d'IA

Développement d'un système d'IA

La base légale : Tout traitement de données personnelles doit reposer sur une des bases légales prévues par les textes. Dans le cas d'un système d'IA développé pour un projet de recherche, la base légale pouvant être utilisée est l'exécution de la *mission d'intérêt public* d'Inria. Dans les autres cas il est possible d'utiliser *l'intérêt légitime*. Il faut alors prouver que les intérêts poursuivis par Inria ne créent pas de déséquilibre au détriment des droits et intérêts des personnes dont les données sont traitées.

La finalité : L'objectif poursuivi par l'utilisation des données doit être établi connu et compatible avec les missions d'Inria. Certains systèmes d'IA sont développés pour un usage opérationnel précis, dans ce cas la finalité est directement liée à celle du traitement futur. Il est tout à fait possible que le développement du système d'IA corresponde à plusieurs usages et donc qu'il ait plusieurs finalités.

La minimisation : Seules les données nécessaires à la finalité du traitement peuvent être utilisées. Mais cela n'empêche pas d'entraîner un algorithme avec beaucoup de données, son application impose de sélectionner les données nécessaires au système d'entraînement (catégorie, typologie, sources, sensibilité des données) puis de mettre en œuvre les moyens techniques pour n'utiliser que celles-ci.

La durée de conservation des données d'entraînement : Pour le développement d'un système d'IA, le fournisseur doit fixer et justifier en amont la durée de conservation des données utilisées. Cette durée doit correspondre à la durée depuis la collecte des données puis l'entraînement du modèle jusqu'à la phase de test déterminant les caractéristiques et performances du produit fini. A l'issue de cette durée, il est possible de conserver ces données pour des opérations de maintenance ou d'amélioration du système. Mais il est toutefois nécessaire d'appliquer le même principe de minimisation à ces étapes donc il est important de ne conserver et utiliser uniquement les données nécessaires à ces dernières.

La réutilisation de bases de données, notamment accessibles en ligne : Pour la constitution d'une base de données à des fins d'entraînement plusieurs cas sont envisageables :

- Inria réutilise des données qu'il a collectées pour d'autres finalités. En principe il faut déterminer si cette réutilisation est compatible avec le traitement initial (il s'agit d'un « test de compatibilité »). Ce test n'est pas nécessaire lorsque la réutilisation est effectuée dans le cadre de la recherche scientifique. Dans ce cas, le traitement est présumé compatible s'il respecte le RGPD et qu'il n'est pas utilisé pour prendre des décisions à l'égard des personnes ;
- Inria utilise une base de données accessible publiquement : Inria doit s'assurer que la constitution de ce jeu de données n'est pas manifestement illicite, par exemple grâce à la mention des sources ;
- Inria réutilise des données qui sont partagés par un tiers : comme dans le cas de données librement accessibles, Inria doit s'assurer qu'il n'est pas en train de réutiliser une base dont la constitution ou le partage était manifestement illicite. Il est important de contractualiser avec ce tiers.

L'information des personnes dont les données personnelles servent à l'entraînement d'un modèle d'IA et sont potentiellement mémorisées par celui-ci : il s'agit du principe de transparence qui s'applique à tout traitement de données personnelles. La difficulté se pose lorsque les données d'entraînement n'ont pas été collectées directement auprès des personnes concernées, notamment

via des API ou des outils de moissonnage.En principe il est obligatoire d'informer individuellement les personnes concernées mais des dérogations peuvent être appliquées :

- Les personnes ont déjà été informées de ce traitement lors de la création du jeu de données (finalités, responsables de traitement, etc.)
- L'information exige des efforts disproportionnés au regard des risques pour les personnes concernées. Dans ce cas il faut réaliser une analyse au cas par cas en tenant compte de la présence ou pas de moyens de contact, de la sensibilité des données, de type de personnes concernées, de la pseudonymisation des données, etc. Cette dérogation s'appliquera facilement à Inria lors de la constitution des bases de données d'apprentissage à des fins de recherche scientifique et notamment lors de la réutilisation des données licites accessibles publiquement qui ne présentent pas de risques pour les personnes concernées.

Dans le cas des dérogations à l'information individuelle une information collective doit être effectuée, notamment sur le site du projet.

Droits d'accès, de rectification, d'opposition et d'effacement des données personnelles. La protection de la vie privée doit être prévue dès la conception du modèle avec une attention particulière pour les données personnelles présentes dans les bases d'entraînement :

- En s'efforçant de rendre les modèles anonymes, lorsque cela n'est pas contraire à l'objectif poursuivi ;
- Et en développant des solutions innovantes pour empêcher la divulgation de données personnelles confidentielles par le modèle.

Le coût, l'impossibilité ou les difficultés pratiques pourront parfois justifier un refus d'exercice des droits. Il est important pour les développeurs ou déployeurs de système d'IA de se tenir informés de l'évolution de l'état de l'art pour protéger au mieux les droits des personnes.

Niveaux de responsabilité. Le responsable du traitement au sens du RGPD est la personne morale qui détermine les moyens et les finalités du traitement envisagé.

Lors du développement d'un système d'IA, le fournisseur qui est à l'initiative de ce dernier et qui en constitue la base de données d'apprentissage à partir de données qu'il a sélectionnées est le responsable de traitement.

Lors de la réutilisation d'un jeu de données déjà existant, le diffuseur et le ré-utilisateur des données seront chacun responsable de leurs propres traitements.

Le sous-traitant traite les données pour le compte du responsable du traitement. Ce dernier doit s'assurer que le sous-traitant respecte le RGPD et qu'il ne traite les données que sur ses instructions dans le cadre d'un contrat de sous-traitance. Un fournisseur de système d'IA peut être sous-traitant s'il développe un système pour le compte d'un de ses clients.

Plusieurs responsables de traitement qui constituent la base de données d'apprentissage d'un système d'IA pour un objectif défini ensemble, peuvent être qualifiés de responsables conjoints du traitement.

Utilisation d'un système d'IA

Le traitement de données personnelles réalisé via le système d'IA doit être mis en conformité aux principes du RGPD. L'utilisateur du système d'IA est responsable du traitement de données personnelles réalisé et le système d'IA peut avoir un rôle de sous-traitant.

Le niveau de sensibilité des données doit être évalué afin de déterminer si le système d'IA est adapté au traitement. Le dépôt de données personnelles dans le système d'IA peut être considéré comme un transfert de données personnelles qui est soumis au RGPD.

Annexe 4 – AI Act et IA générative

Le règlement européen sur l'intelligence artificielle (AI Act) adopté en 2024 établit un cadre juridique encadrant le développement, l'usage et la mise sur le marché des systèmes d'IA, y compris l'IA générative. Les principales dispositions concernant l'IA générative sont :

- Classification des systèmes d'IA générative comme « à usage général » plutôt que « à haut risque » sauf application spécifique (ex. santé, justice).
- Exigences de transparence pour les systèmes d'IA générant ou manipulant des contenus audio, vidéo ou textuels.
- Obligation de divulgation : le contenu généré par une IA doit être marqué et signalé aux utilisateurs.
- Mise en place de mesures pour prévenir la génération de contenus illégaux, notamment via le filtrage et la supervision humaine.

Exigences en matière de transparence et traçabilité

L'AI Act impose des obligations de transparence aux développeurs et utilisateurs d'IA générative, afin de garantir une utilisation responsable et traçable de l'IA (article 50) :

Pour les concepteurs de modèles d'IA :

- Veiller à ce que les systèmes d'IA destinés à interagir directement avec des personnes physiques soient conçus et développés de manière à ce que les personnes physiques concernées soient informées qu'elles interagissent avec un système d'IA (à moins que cela ne soit évident). Exception pour les systèmes d'IA autorisés par la loi à détecter, prévenir, enquêter ou poursuivre des infractions pénales (sous conditions).
- Documentation détaillée des données d'entraînement utilisées.
- Intégration de filigranes (watermarking) et métadonnées pour permettre l'identification des contenus générés. Cette obligation ne s'applique pas dans la mesure où les systèmes d'IA remplissent une fonction d'assistance pour l'édition standard ou ne modifient pas substantiellement les données d'entrée fournies par le déployeur ou leur sémantique, ou lorsqu'ils sont autorisés par la loi à détecter, prévenir, enquêter ou poursuivre des infractions pénales.
- Traçabilité des sorties générées pour garantir une vérifiabilité et prévenir les abus.
- Publication de résumés sur le fonctionnement des modèles afin d'assurer une meilleure explicabilité.

Pour les utilisateurs, déployeurs et exploitants de l'IA :

- Obligation de mentionner explicitement qu'un contenu a été généré ou modifié par une IA.
- Obligation d'indiquer les manipulations effectuées, notamment pour les contenus diffusés dans un cadre informatif ou audio-visuel.

Interaction avec d'autres réglementations européennes

L'AI Act interagit avec plusieurs autres régulations européennes, notamment :

Règlement Général sur la Protection des Données (RGPD - 2016/679) :

- Encadrement de l'usage des données personnelles dans l'entraînement des modèles d'IA.

- Mesures de minimisation des données et anonymisation pour limiter les risques de réidentification.

Directive Copyright 2019/790 (Text and Data Mining - TDM) :

- Réglementation sur l'utilisation d'œuvres protégées par le droit d'auteur pour l'entraînement des modèles.
- Exception prévue pour les instituts de recherche et établissements publics sous certaines conditions.

Règlement sur la gouvernance des données (Data Governance Act - 2022/868) :

- Favorise le partage de données dans un cadre sécurisé pour les projets d'IA.
- Encourage le développement de bases de données ouvertes à des fins scientifiques et d'innovation.

Application aux organismes publics de recherche

Les organismes publics de recherche bénéficient de certaines dérogations et flexibilités dans l'AI Act afin de préserver la liberté académique et favoriser l'innovation.

- Exemptions pour les activités de recherche (article 25 de l'AI Act). *Extrait : « Le présent règlement devrait soutenir l'innovation et respecter la liberté scientifique et ne devrait pas compromettre les activités de recherche et de développement.* ***Il est donc nécessaire d'exclure de son champ d'application les systèmes et modèles d'IA spécifiquement développés et mis en service aux seules fins de la recherche et du développement scientifiques.*** *En outre, il est nécessaire de veiller à ce que le présent règlement* ***n'affecte pas autrement les activités de recherche et de développement scientifiques relatives aux systèmes ou modèles d'IA avant leur mise sur le marché ou leur mise en service.*** *En ce qui concerne les activités de recherche, d'essai et de développement axées sur les produits, relatives aux systèmes ou modèles d'IA, les dispositions du présent règlement ne devraient pas non plus s'appliquer avant la mise en service ou la mise sur le marché de ces systèmes et modèles****. Cette exclusion est sans préjudice de l'obligation de se conformer au présent règlement lorsqu'un système d'IA relevant du champ d'application du présent règlement est mis sur le marché ou mis en service à la suite de cette activité de recherche et de développement, et sans préjudice de l'application des dispositions relatives aux bacs à sable réglementaires de l'IA et aux essais en conditions réelles.*** *En outre, sans préjudice de l'exclusion des systèmes d'IA spécifiquement développés et mis en service aux seules fins de la recherche et du développement scientifique, tout autre système d'IA susceptible d'être utilisé pour mener une activité de recherche et de développement devrait rester soumis aux dispositions du présent règlement. En tout état de cause, toute activité de recherche et de développement devrait être menée conformément à des normes éthiques et professionnelles reconnues en matière de recherche scientifique et dans le respect du droit de l'Union applicable. »*
- Possibilité d'utiliser des données protégées pour l'entraînement des modèles à des fins de recherche (directive européenne 2016 Copyright en matière de *Text and Data Mining* (TDM) + Directive 2019/790)
- Obligations allégées en matière de documentation pour les prototypes de recherche : documentation simplifiée pour les modèles non encore commercialisés et absence d'obligation de certification tant que le système reste en phase expérimentale.

Annexe 5 - Abréviations

- CORUM : Comité de réflexion sur la responsabilité en recherche en sciences du numérique
- DAFP : Direction des affaires financières et patrimoniales
- DAJ : Direction des affaires juridiques
- DPD : Délégué à la Protection des Données
- DGDI : Direction générale déléguée à l'innovation
- DGDS : Direction générale déléguée à la science
- FSD : Fonctionnaire sécurité défense
- PPST : protection du potentiel scientifique et technique de la Nation
- RGPD : Règlement général de protection des données
- RSSI : Responsable de la sécurité des systèmes d'information
- SSI : sécurité des systèmes d'information