AYI NEDJIMI Consultants

IA et Conformité RGPD : Données Personnelles dans les

Catégorie : Intelligence Artificielle Lecture : 26 min Publié le : 13/02/2026 Auteur : Ayi NEDJIMI

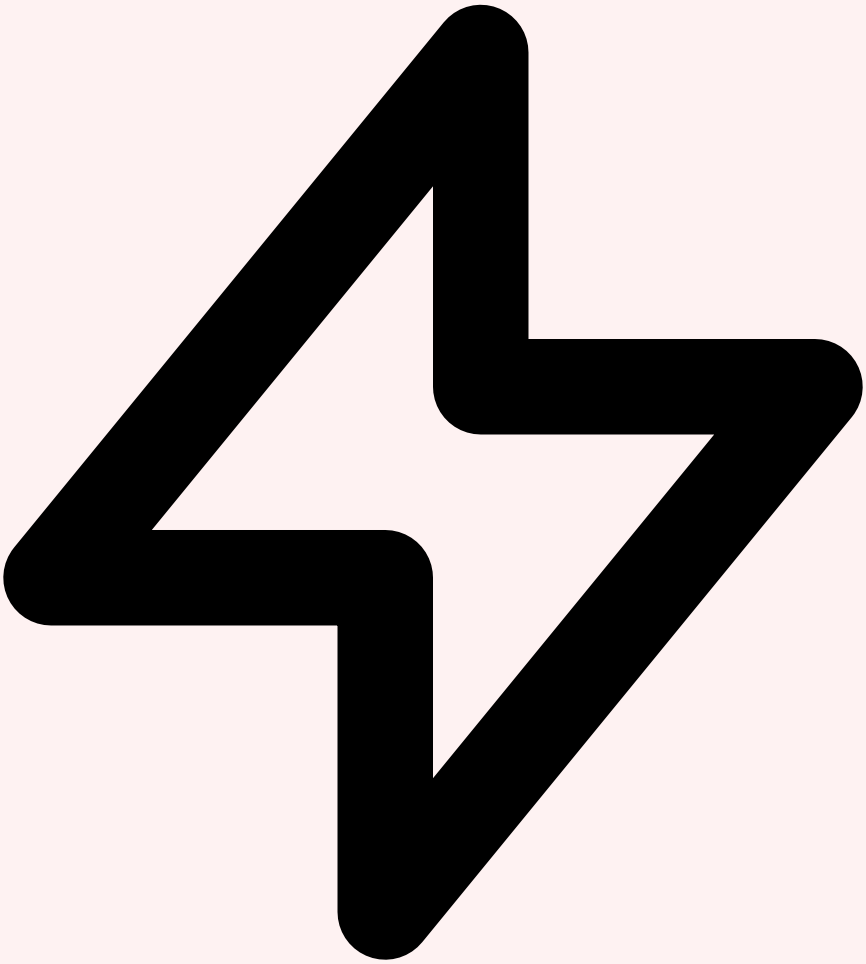
Guide complet sur la conformité RGPD pour l'IA : base légale du traitement, minimisation des données, droit à l'oubli dans les LLM, DPIA,. Guide.

Table des Matières

Avez-vous évalué les risques d'injection de prompt sur vos systèmes d'IA en production ?

1 Le RGPD Face au Défi de l'IA Générative

Le **Règlement Général sur la Protection des Données (RGPD)**, entré en vigueur le 25 mai 2018, constitue le socle juridique européen en matière de protection des données personnelles. Conçu dans un contexte technologique où les traitements de données étaient relativement prévisibles et déterministes, ce règlement se retrouve aujourd'hui confronté à un **défi majeur** : l'émergence de l'intelligence artificielle générative et des grands modèles de langage (LLM). En 2026, la tension entre les principes fondamentaux du RGPD et le fonctionnement intrinsèque des systèmes d'IA est devenue l'un des enjeux juridiques et techniques les plus complexes du paysage numérique européen. Les entreprises qui déploient, entraînent ou utilisent des modèles d'IA se trouvent dans un labyrinthe réglementaire où chaque décision technique a des implications juridiques potentiellement considérables.



La collision entre deux références

Le RGPD repose sur des principes clairs : **finalité déterminée** du traitement, **minimisation des données**, **limitation de la conservation**, **transparence** et **droits individuels** exercables à tout moment. Or, le fonctionnement des LLM entre en friction directe avec plusieurs de ces principes. L'entraînement d'un modèle comme GPT-4, Claude ou Llama nécessite l'ingestion de quantités massives de données textuelles — souvent des milliards de tokens — parmi lesquelles figurent inévitablement des données personnelles : noms, adresses, numéros de téléphone, informations médicales, opinions politiques ou religieuses. Ces données sont absorbées dans les paramètres du modèle de manière distribuée et non réversible, rendant leur identification, leur extraction et leur suppression extraordinairement difficiles, voire impossibles avec les techniques actuelles. Le principe de finalité est également mis à rude épreuve : un modèle entraîné sur un corpus donné peut ensuite être utilisé pour des finalités très différentes de celles envisagées lors de la collecte initiale des données.

Positions des autorités de protection des données

Les autorités européennes de protection des données ont adopté des positions variées mais convergentes sur la question. La **CNIL française** a publié en 2024 une série de recommandations spécifiques aux systèmes d'IA, précisant que le RGPD s'applique pleinement à toutes les phases du cycle de vie d'un modèle : collecte des données d'entraînement, pré-traitement, entraînement proprement dit, validation, déploiement et inférence. Le **Garante italiano** (autorité italienne) a créé un précédent majeur en mars 2023 en interdisant temporairement ChatGPT sur le territoire italien, invoquant l'absence de base légale pour le traitement massif de données personnelles, le défaut d'information des personnes concernées et l'absence de mécanisme de vérification de l'âge. Cette décision, bien que levée un mois plus tard après qu'OpenAI eut mis en place des mesures correctives, a eu un impact considérable sur l'ensemble de l'écosystème européen de l'IA.

Notre avis d'expert

La gouvernance de l'IA est le prochain grand chantier de la cybersécurité. Les attaques par prompt injection, l'empoisonnement de données d'entraînement et l'extraction de modèles sont des menaces concrètes que nous observons de plus en plus lors de nos missions. Ne pas s'y préparer, c'est accepter un risque majeur.

L'**European Data Protection Board (EDPB)** a renforcé cette approche en créant une task force dédiée à ChatGPT et aux modèles de langage en général, aboutissant à un rapport en décembre 2024 qui établit des lignes directrices harmonisées. Ce rapport souligne que la **légitimité du traitement ne peut être présumée** du seul fait de l'innovation technologique, et que les développeurs et déployeurs de systèmes d'IA doivent démontrer activement leur conformité au RGPD à chaque étape. Le rapport de l'EDPB distingue également clairement les responsabilités entre le **fournisseur du modèle** (qui entraîne le LLM) et le **déployeur** (qui l'intègre dans un service), chacun devant justifier d'une base légale distincte pour son propre traitement.

Point clé : Le RGPD n'interdit pas l'IA — il impose un cadre strict qui nécessite une approche proactive de conformité. Les entreprises qui intègrent la protection des données dès la conception de leurs projets IA (privacy by design) bénéficient d'un avantage concurrentiel significatif en réduisant les risques juridiques et en renforçant la confiance de leurs utilisateurs et clients.

En pratique, l'affaire **OpenAI vs Garante** a établi un précédent juridique déterminant pour l'ensemble du secteur. Les mesures correctives imposées à OpenAI — publication d'une politique de confidentialité détaillée, mise en place d'un mécanisme d'opt-out pour l'entraînement, implémentation de la vérification d'âge et offre d'un droit d'opposition effectif — sont devenues de facto le standard minimum attendu par les autorités européennes pour tout fournisseur de service d'IA. Cette affaire a également démontré que les autorités de protection des données disposent d'outils juridiques puissants pour réguler l'IA, même en l'absence d'un cadre réglementaire spécifique à l'intelligence artificielle, l'AI Act n'étant entré en application que progressivement à partir de 2024.

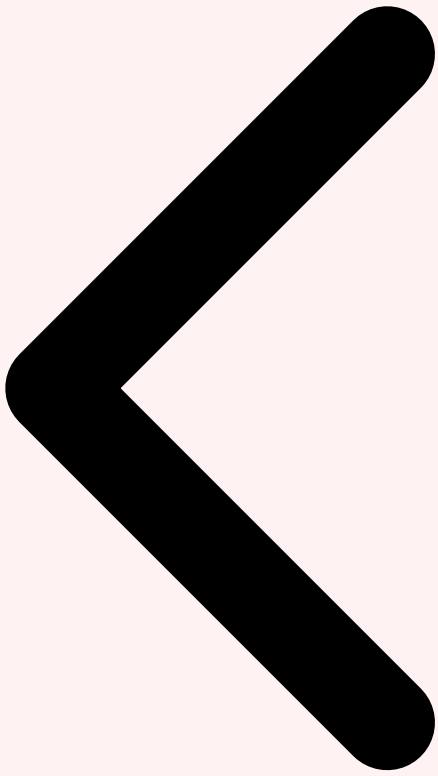
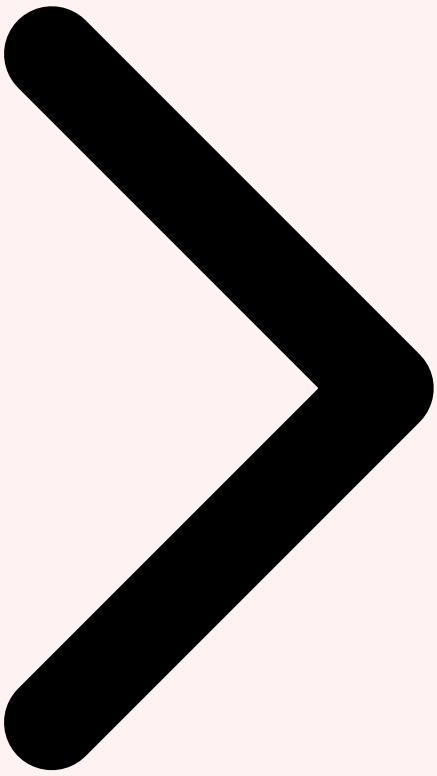


Table des Matières RGPD et Défi IA Base Légale

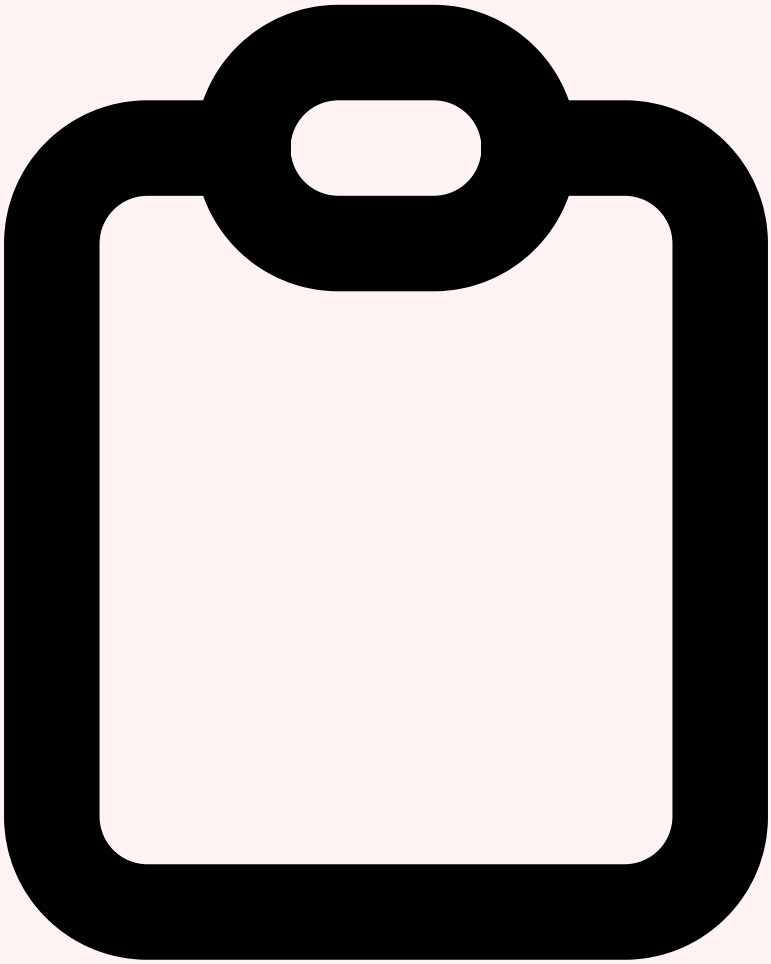


Critere	Description	Niveau de risque
Confidentialite	Protection des donnees d'entrainement et des prompts	Eleve
Integrite	Fiabilite des sorties et detection des hallucinations	Critique
Disponibilite	Resilience du service et gestion de la charge	Moyen
Conformite	Respect du RGPD, AI Act et politiques internes	Eleve

2 Base Légale du Traitement des Données par l'IA

L'**article 6 du RGPD** constitue la pierre angulaire de toute démarche de conformité pour les projets d'IA. Il établit six bases légales possibles pour justifier un traitement de données personnelles, et le choix de la base appropriée conditionne l'ensemble des obligations et des droits qui en découlent. Pour les acteurs de l'IA, cette question revêt une importance capitale car elle détermine non seulement la légalité du traitement mais aussi l'étendue des droits des personnes concernées et les obligations de documentation. En 2026, après plusieurs années de débats entre les autorités de protection des données, les entreprises

technologiques et les juristes spécialisés, un consensus commence à émerger sur les pratiques acceptables, bien que des zones grises subsistent, notamment pour l'entraînement de modèles sur des données collectées à grande échelle sur Internet.



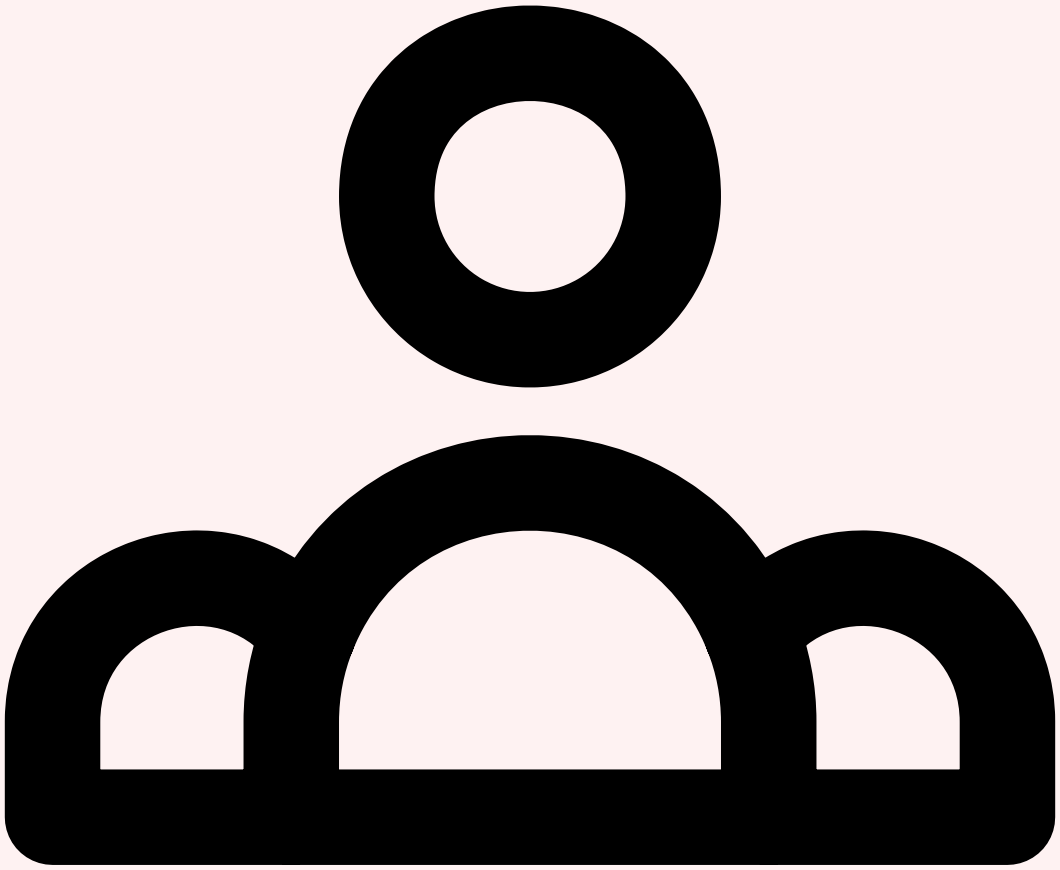
Les six bases légales appliquées à l'IA

Le **consentement** (article 6.1.a) est la base légale la plus intuitive mais aussi la plus contraignante. Pour être valide dans le contexte de l'IA, le consentement doit être libre, spécifique, éclairé et univoque. Cela signifie que l'utilisateur doit comprendre précisément comment ses données seront utilisées pour entraîner ou faire fonctionner un modèle d'IA, et doit pouvoir retirer son consentement à tout moment — avec des conséquences effectives sur le traitement. Dans la pratique, cette base légale est difficile à mettre en oeuvre pour l'entraînement de LLM car le retrait du consentement impliquerait théoriquement de réentraîner le modèle sans les données de la personne concernée, ce qui est techniquement et économiquement irréaliste pour des modèles de fondation.

L'**intérêt légitime** (article 6.1.f) est la base légale la plus fréquemment invoquée par les développeurs de modèles d'IA pour justifier l'entraînement sur des données collectées publiquement. Cette base exige une mise en balance entre les intérêts de l'organisation qui traite les données et les droits et libertés des personnes concernées. OpenAI, Anthropic, Google et d'autres grands acteurs se fondent sur cette base en argumentant que le développement de l'IA constitue un intérêt légitime contribuant au progrès technologique et économique, tout en mettant en place des mesures de sauvegarde (anonymisation partielle, filtrage, opt-out). La CNIL a néanmoins rappelé que l'intérêt légitime nécessite une analyse rigoureuse, documentée dans un **test de proportionnalité** (Legitimate Interest Assessment — LIA), et que les droits des personnes — notamment le droit d'opposition — doivent être effectivement respectés.

Cas concret
L'attaque par prompt injection sur les systèmes GPT documentée par OWASP en 2023 a révélé que des instructions malveillantes dissimulées dans des documents pouvaient détourner le comportement de chatbots d'entreprise, accédant à des données internes sensibles sans aucune authentification supplémentaire.

Vos pipelines de données d'entraînement sont-ils protégés contre l'empoisonnement ?

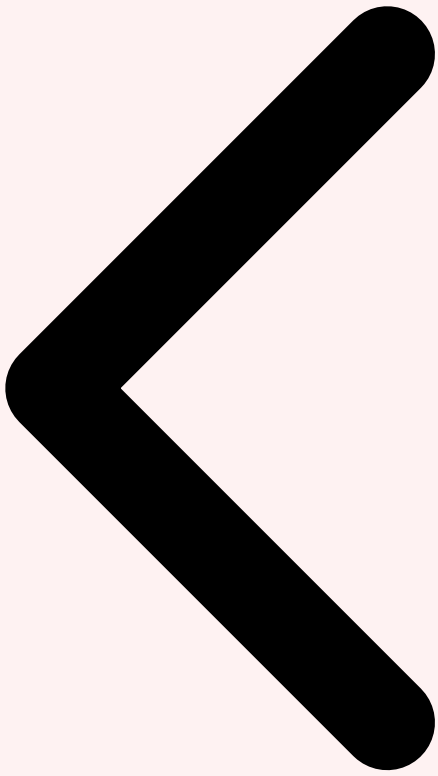


Contrat, obligation légale et intérêt public

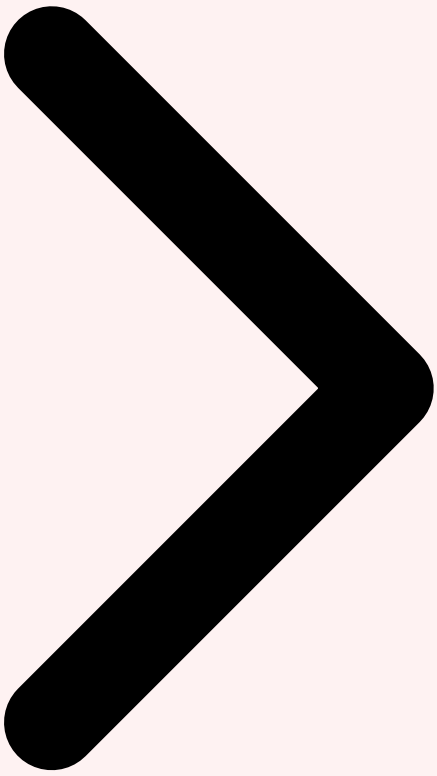
L'**exécution d'un contrat** (article 6.1.b) constitue une base légale pertinente pour les services IA B2B et les applications d'IA intégrées à un service existant. Lorsqu'un utilisateur souscrit à un service incluant explicitement des fonctionnalités IA — par exemple un assistant de rédaction, un outil de traduction automatique ou un système de recommandation —, le traitement des données nécessaire à la fourniture de ce service peut être fondé sur la nécessité contractuelle. Toutefois, cette base ne couvre que les traitements strictement nécessaires à l'exécution du contrat, et non les traitements annexes comme l'utilisation des données d'interaction pour améliorer ou réentraîner le modèle. Le fine-tuning d'un modèle sur les données d'un client B2B requiert soit un consentement distinct, soit une justification par l'intérêt légitime avec les garanties appropriées.

L'**obligation légale** (article 6.1.c) et la **mission d'intérêt public** (article 6.1.e) sont des bases légales mobilisées par les acteurs publics et les organisations soumises à des obligations réglementaires spécifiques. Les autorités fiscales, les organismes de santé publique ou les services de sécurité nationale peuvent fonder leurs traitements IA sur ces bases lorsque la loi les y autorise ou les y oblige. Par exemple, l'utilisation de l'IA pour la

détection de fraude fiscale à grande échelle relève de la mission d'intérêt public, tandis que le traitement automatisé de données de santé pour la recherche épidémiologique peut s'appuyer sur l'obligation légale issue des codes de santé publique. Enfin, la **sauvegarde des intérêts vitaux** (article 6.1.d) reste marginale dans le contexte de l'IA, potentiellement applicable à des systèmes de diagnostic médical d'urgence ou d'alerte de catastrophe naturelle.

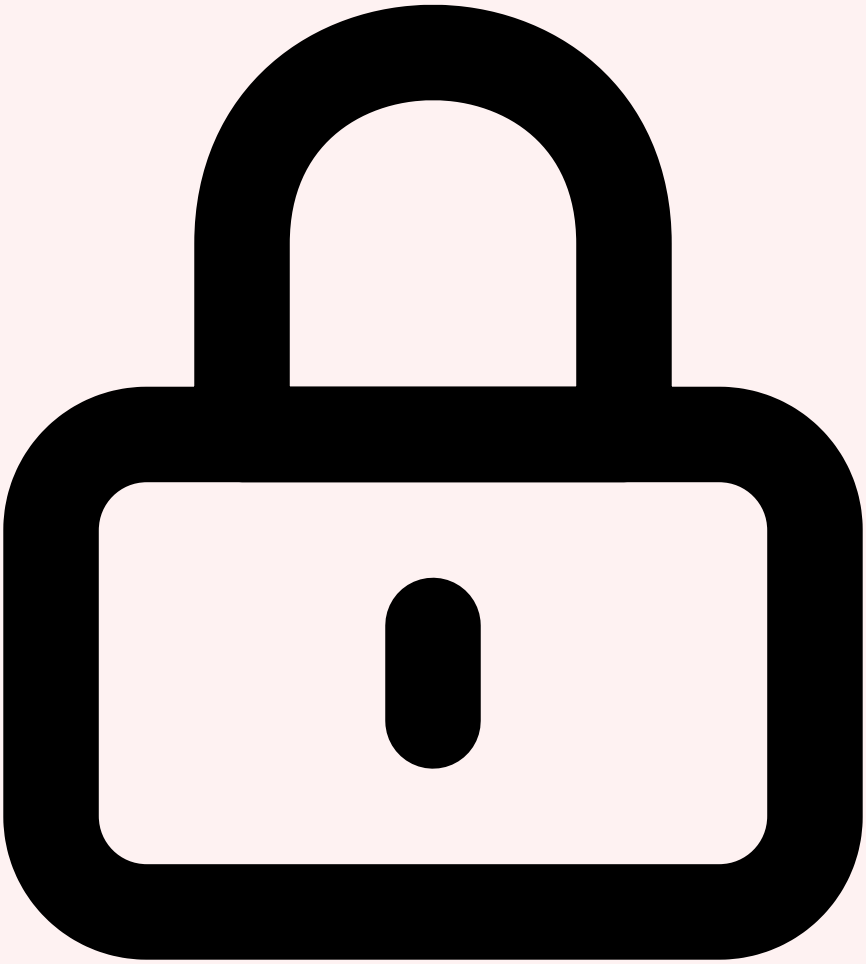


RGPD et Défi IA Base Légale Minimisation et Privacy



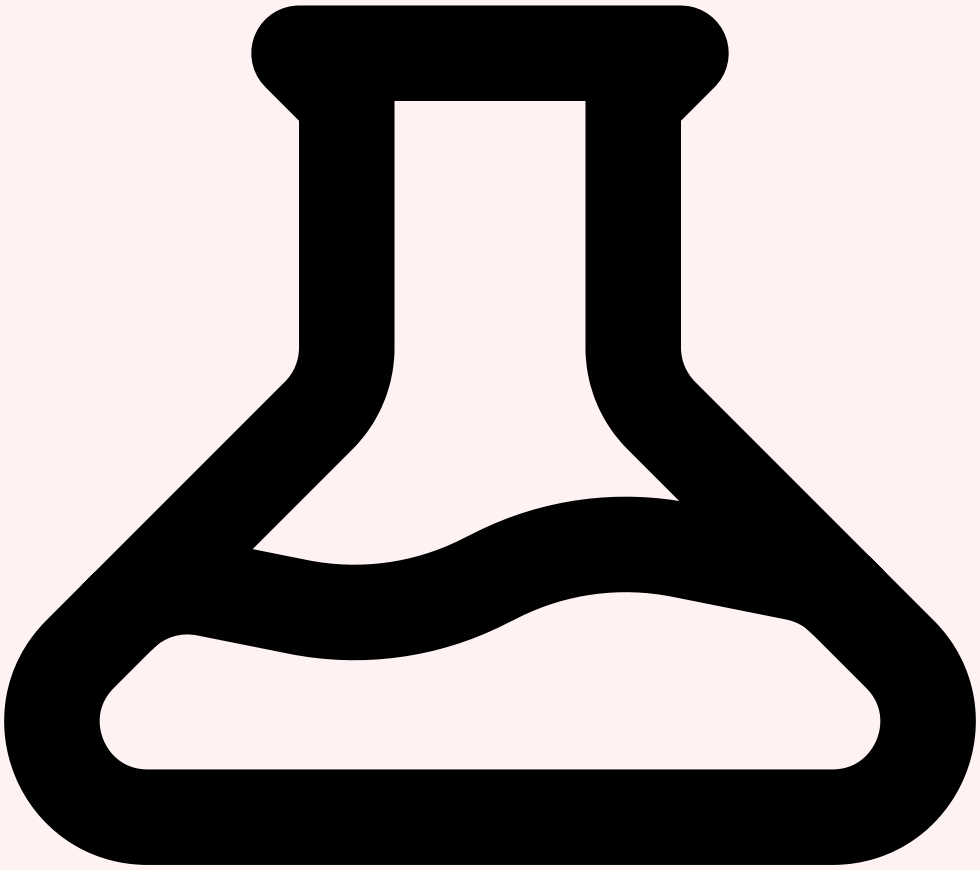
3 Minimisation des Données et Privacy by Design

Le principe de **minimisation des données** (article 5.1.c du RGPD) impose que les données personnelles collectées et traitées soient adéquates, pertinentes et limitées à ce qui est nécessaire au regard des finalités du traitement. Appliqué aux systèmes d'IA, et en particulier aux LLM, ce principe constitue un défi technique majeur. L'entraînement d'un modèle de langage performant nécessite par nature des volumes considérables de données textuelles, et la tentation est grande de maximiser la quantité et la diversité des données d'entraînement pour améliorer les performances du modèle. Pourtant, le RGPD exige que les développeurs démontrent que chaque catégorie de données utilisée est strictement nécessaire et qu'aucune alternative moins intrusive n'existe pour atteindre le même objectif. Cette exigence impose une discipline rigoureuse dans la sélection, le filtrage et le pré-traitement des données d'entraînement.



Privacy by Design pour les projets LLM (article 25)

L'**article 25 du RGPD** consacre le principe de protection des données dès la conception (privacy by design) et par défaut (privacy by default). Pour les projets LLM, cela signifie que la protection des données personnelles doit être intégrée à chaque étape du développement, depuis la conception de l'architecture du modèle jusqu'à son déploiement en production. En pratique, le privacy by design se traduit par plusieurs mesures concrètes : la mise en œuvre de pipelines de nettoyage et de filtrage des données d'entraînement pour éliminer les données personnelles identifiables avant l'ingestion par le modèle, l'implémentation de mécanismes de détection et de suppression des PII (Personally Identifiable Information) dans les données entrantes, la configuration par défaut des services pour minimiser la collecte (pas de journalisation des prompts, pas de rétention des conversations sauf opt-in explicite), et l'architecture de systèmes de guardrails empêchant le modèle de divulguer des informations personnelles apprises lors de l'entraînement.

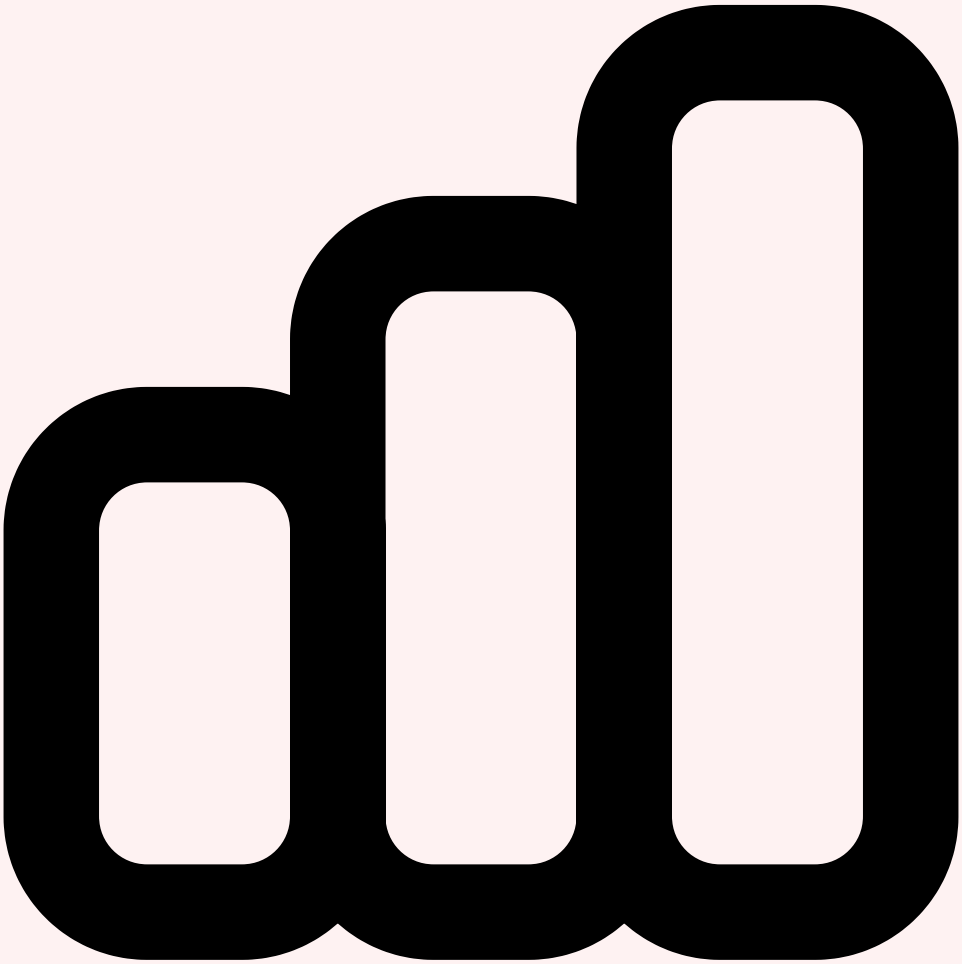


Techniques d'anonymisation et de pseudonymisation

L'**anonymisation** consiste à transformer irréversiblement les données personnelles de sorte qu'aucune ré-identification ne soit possible, même en combinant les données avec d'autres sources. Les données véritablement anonymisées sortent du champ d'application du RGPD, ce qui en fait la solution idéale pour l'entraînement de modèles. Toutefois, l'anonymisation parfaite est extraordinairement difficile à atteindre dans le contexte des données textuelles, car les informations personnelles sont souvent imbriquées dans le contexte sémantique du texte. Un simple remplacement des noms propres par des tokens génériques peut être insuffisant si le contexte permet la ré-identification par recoupement. La **pseudonymisation**, quant à elle, consiste à remplacer les identifiants directs par des pseudonymes, tout en conservant la possibilité de relier les données à la personne via une table de correspondance séparée. Elle réduit les risques mais ne dispense pas de l'application du RGPD, car les données restent considérées comme personnelles.

Les **données synthétiques** (synthetic data) représentent une approche prometteuse pour résoudre le dilemme entre performance des modèles et protection de la vie privée. Plutôt que d'utiliser des données personnelles réelles, il est possible de générer des données

artificielles qui préservent les propriétés statistiques et les patterns linguistiques des données originales sans contenir d'informations identifiantes. Des entreprises comme Mostly AI, Gretel et Tonic.ai proposent des solutions de génération de données synthétiques certifiées RGPD-compatibles. L'entraînement ou le fine-tuning d'un modèle sur des données synthétiques élimine le risque de mémorisation de données personnelles réelles. Cependant, la qualité et la représentativité des données synthétiques restent des sujets de recherche actifs, et un modèle entraîné exclusivement sur des données synthétiques peut présenter des biais ou des lacunes par rapport à un modèle entraîné sur des données réelles.

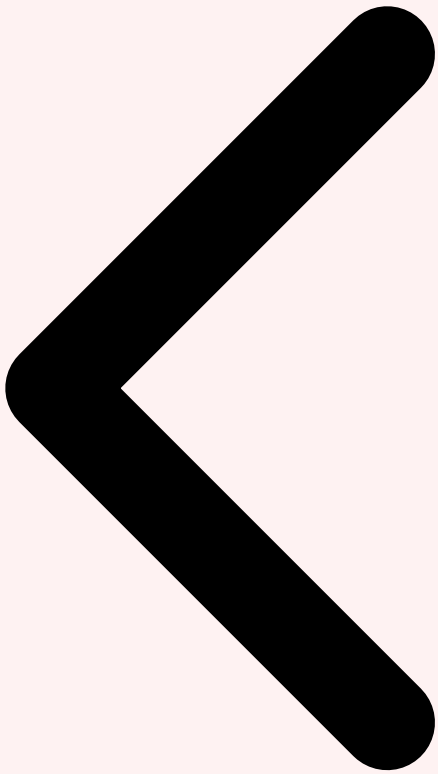


Differential Privacy et politiques de rétention

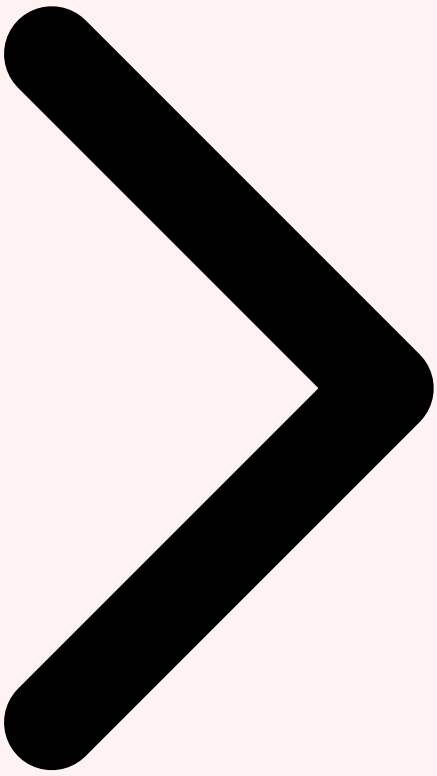
La **confidentialité différentielle** (differential privacy) est une technique mathématiquement fondée qui ajoute du bruit calibré aux données ou aux gradients pendant l'entraînement du modèle, garantissant qu'aucune donnée individuelle ne puisse être extraite ou inférée à partir du modèle final. Google a été un pionnier de cette approche avec son framework DP-SGD (Differentially Private Stochastic Gradient Descent), et Apple l'utilise depuis plusieurs années dans ses produits. Dans le contexte des LLM, la differential privacy peut être appliquée au fine-tuning (DP-LoRA, par exemple) avec un

compromis mesurable entre le niveau de protection (paramètre epsilon) et la qualité du modèle résultant. Un epsilon faible offre une protection forte mais peut dégrader significativement les performances du modèle, tandis qu'un epsilon élevé préserve les performances mais offre une protection plus limitée. La recherche en 2026 progresse rapidement vers des méthodes offrant un meilleur compromis protection-performance, notamment via des techniques de sous-échantillonnage intelligent et de composition adaptative du bruit.

Les **politiques de rétention des données** constituent un aspect souvent négligé mais crucial de la conformité RGPD pour les services d'IA. L'article 5.1.e du RGPD impose que les données personnelles soient conservées sous une forme permettant l'identification des personnes concernées pendant une durée n'excédant pas celle nécessaire au regard des finalités du traitement. Pour les services d'IA conversationnelle, cela implique de définir des durées de conservation claires pour les données de prompts et les conversations des utilisateurs. OpenAI conserve les conversations 30 jours par défaut (avec possibilité de désactivation), Anthropic propose une politique similaire, et les solutions on-premise comme Ollama ou vLLM permettent une rétention zéro. Les entreprises déployant des solutions d'IA doivent documenter leurs politiques de rétention, implémenter des mécanismes de suppression automatique à expiration, et s'assurer que les données de prompts ne sont pas réutilisées pour l'entraînement sans base légale appropriée.



Base Légale Minimisation et Privacy Droit à l'Oubli LLM



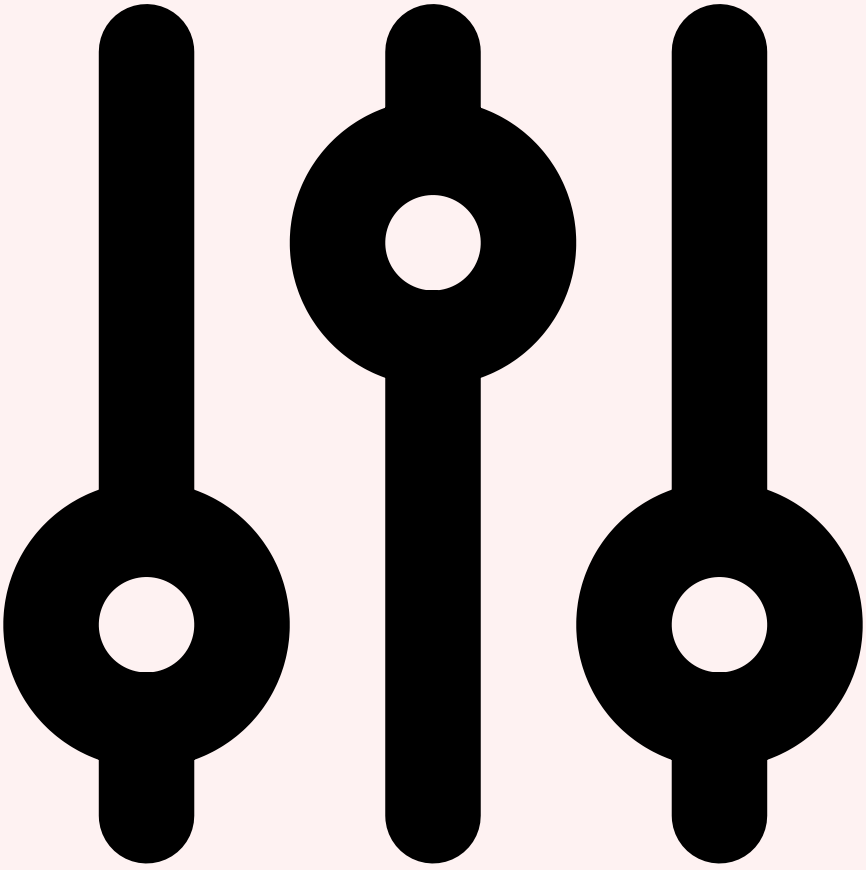
4 Droit à l'Oubli et LLM : Le Défi Technique

L'**article 17 du RGPD** consacre le droit à l'effacement, communément appelé « droit à l'oubli », permettant à toute personne de demander la suppression de ses données personnelles lorsque certaines conditions sont remplies : les données ne sont plus nécessaires aux finalités du traitement, le consentement est retiré, la personne exerce son droit d'opposition, les données ont fait l'objet d'un traitement illicite, ou l'effacement est requis par une obligation légale. Ce droit, relativement simple à mettre en oeuvre dans les systèmes de bases de données classiques où les données sont stockées de manière structurée et localisable, se heurte à un obstacle technique fondamental dans le cas des LLM : les données personnelles utilisées pour l'entraînement sont **absorbées et distribuées dans les milliards de paramètres du modèle** de manière non réversible. Il n'existe pas de mécanisme simple pour localiser et supprimer une information spécifique apprise par le modèle sans compromettre l'intégrité et les performances de l'ensemble.

Machine Unlearning : l'état de l'art

Le **machine unlearning** (désapprentissage automatique) est un domaine de recherche en pleine expansion qui vise à développer des techniques permettant de retirer l'influence de données spécifiques d'un modèle déjà entraîné, sans nécessiter un réentraînement complet. Plusieurs approches sont explorées par la communauté scientifique en 2026. Le **réentraînement partiel** (SISA — Sharded, Isolated, Sliced, and Aggregated) consiste à entraîner le modèle sur des sous-ensembles de données indépendants, de sorte que la suppression d'un point de données ne nécessite que le réentraînement du sous-ensemble concerné. Bien que théoriquement élégante, cette approche est impraticable pour les grands modèles de fondation en raison de son coût computationnel et de la perte de cohérence globale. Le **gradient ascent ciblé** vise à inverser l'effet de l'entraînement sur des données spécifiques en appliquant des mises à jour de gradient dans la direction opposée, mais cette technique est instable et peut provoquer des dégradations en cascade des performances du modèle sur d'autres tâches. Pour approfondir, consultez IA pour la Génération de Code : Copilot, Cursor, Claude Code.

Des approches plus récentes et prometteuses incluent le **knowledge editing** (édition de connaissances), qui modifie chirurgicalement les poids du modèle pour altérer ou supprimer des faits spécifiques sans affecter le reste des connaissances. Des techniques comme ROME (Rank-One Model Editing) et MEMIT (Mass-Editing Memory In a Transformer) permettent de localiser et modifier les neurones responsables du stockage d'informations factuelles spécifiques. Toutefois, ces techniques sont encore à un stade expérimental et leur efficacité sur des informations personnelles dispersées dans le modèle (par opposition à des faits factuels localisés) reste limitée. Le **task arithmetic** est une autre approche émergente qui soustrait les vecteurs de tâches correspondant aux données à oublier, offrant un compromis intéressant entre efficacité et préservation des performances globales.



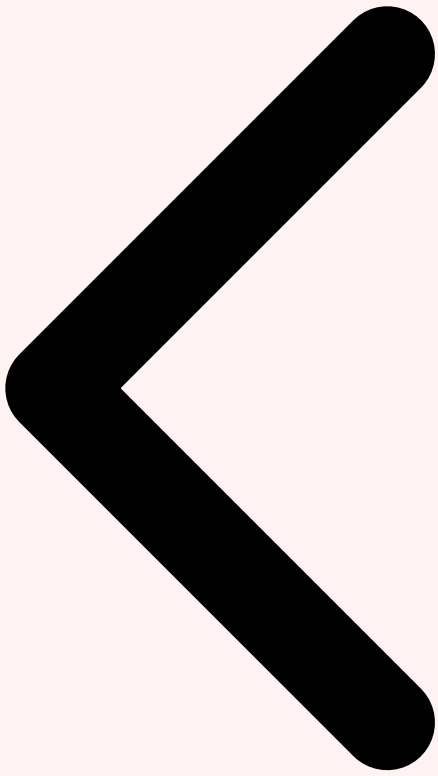
Alternatives pragmatiques et positions des DPAs

Face aux limitations techniques du machine unlearning, des **alternatives pragmatiques** ont émergé pour répondre aux demandes d'effacement dans le contexte des LLM. L'**output filtering** consiste à implémenter des guardrails au niveau de la couche d'inférence pour

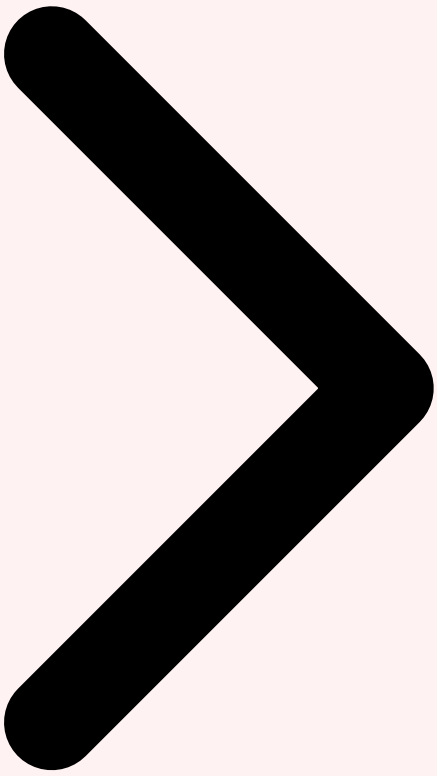
empêcher le modèle de divulguer des informations personnelles spécifiques, même s'il les a mémorisées. Cette approche ne supprime pas les données du modèle mais en bloque la restitution, ce qui pose la question de savoir si elle satisfait réellement l'exigence d'effacement du RGPD. Le **fine-tuning correctif** utilise des techniques d'alignement (RLHF, DPO) pour entraîner le modèle à refuser de restituer certaines informations personnelles, créant une couche de « censure » comportementale. Cette méthode est plus robuste que le simple filtrage d'output mais ne garantit pas que l'information ne puisse être extraite par des techniques de prompt injection avancées ou des attaques adversariales avancées.

Les **autorités de protection des données (DPAs)** européennes ont adopté des positions nuancées sur le droit à l'oubli dans les LLM. La CNIL, dans ses recommandations de 2024, reconnaît que l'effacement complet des données d'un modèle entraîné peut s'avérer disproportionné et techniquement irréalisable, et admet que des mesures alternatives — comme le filtrage des outputs, le blocage de la régurgitation et la mise à jour des données d'entraînement pour les prochaines versions du modèle — peuvent constituer une réponse acceptable, à condition que ces mesures soient documentées et effectivement efficaces. Le Garante italiano a adopté une position plus stricte, exigeant d'OpenAI la mise en œuvre d'un mécanisme permettant aux résidents italiens de demander la correction ou la suppression de données inexactes les concernant générées par ChatGPT. L'EDPB a souligné dans son rapport que la question du droit à l'effacement dans les modèles d'IA nécessite une approche au cas par cas, en distinguant les données présentes dans les données d'entraînement (input data) et les données générées par le modèle (output data).

Recommandations pratiques : Les entreprises utilisant des LLM doivent installer un processus clair de gestion des demandes d'effacement comprenant : (1) un formulaire de demande accessible, (2) un workflow de vérification d'identité, (3) une évaluation technique de la faisabilité, (4) l'implémentation de mesures d'atténuation (output filtering, correction des données), (5) une réponse documentée à la personne concernée dans le délai d'un mois, et (6) la suppression des données des datasets d'entraînement pour les futures versions du modèle.

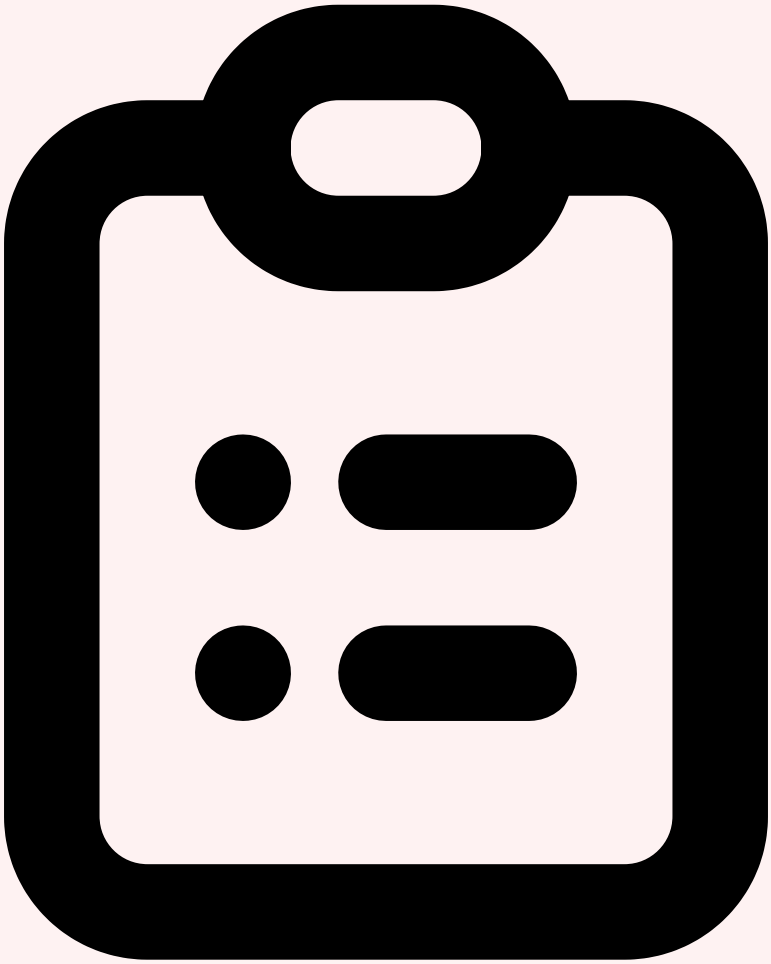


Minimisation et Privacy Droit à l'Oubli LLM DPIA Projets IA



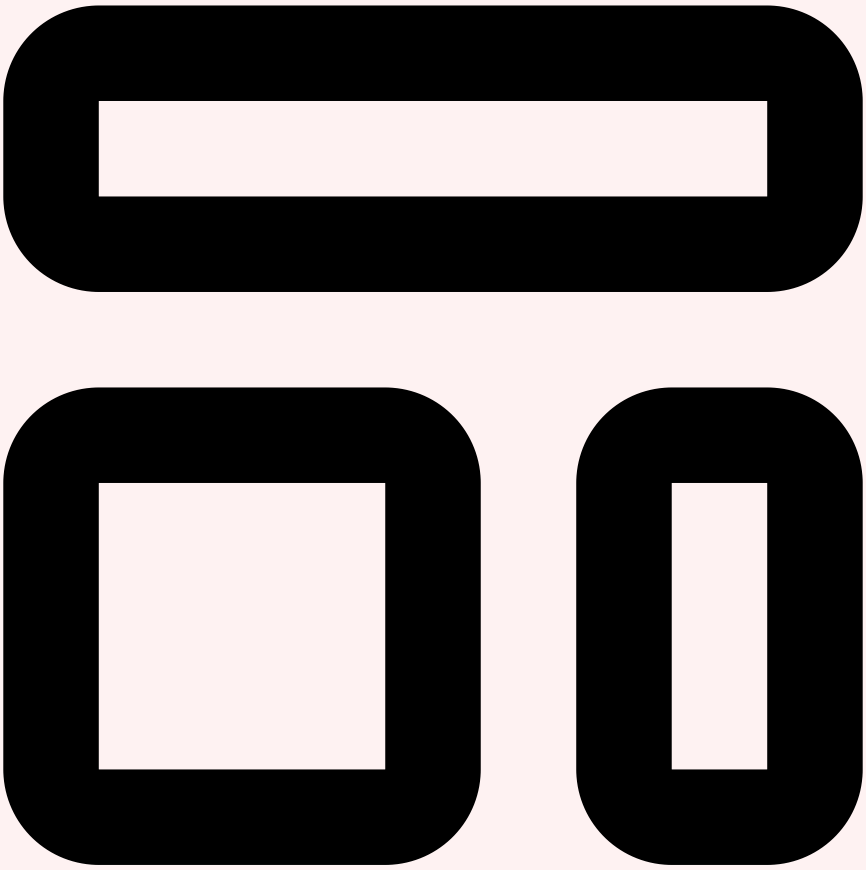
5 DPIA pour les Projets IA

L'**analyse d'impact relative à la protection des données (DPIA)**, prévue par l'article 35 du RGPD, est un exercice d'évaluation des risques obligatoire lorsqu'un traitement est susceptible d'engendrer un risque élevé pour les droits et libertés des personnes physiques. Dans le contexte de l'IA, la quasi-totalité des projets impliquant des données personnelles nécessite une DPIA, compte tenu de la nature systématique et automatisée des traitements, de l'évaluation ou du scoring de personnes, du traitement à grande échelle et de l'utilisation de technologies innovantes — autant de critères identifiés par le Groupe de travail Article 29 (devenu EDPB) comme déclencheurs d'une DPIA. La CNIL a confirmé cette position dans ses recommandations spécifiques IA, précisant qu'une DPIA est requise dès lors qu'un système d'IA traite des données personnelles de manière automatisée pour produire des résultats ayant un effet sur les personnes concernées.



Méthodologie DPIA adaptée aux projets LLM

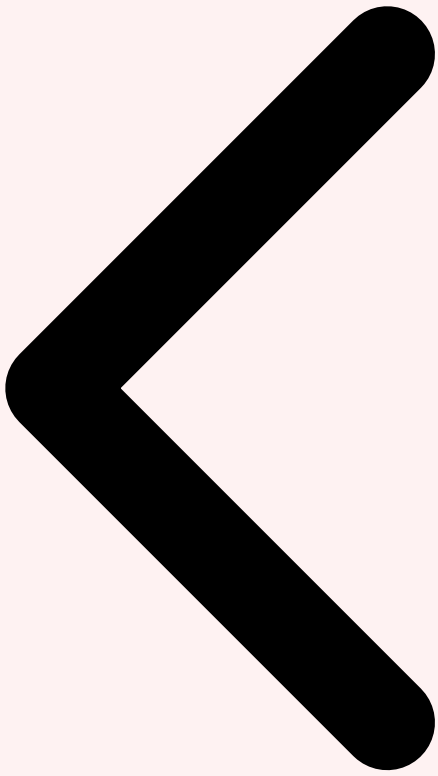
La méthodologie DPIA pour un projet LLM doit couvrir l'ensemble du cycle de vie du système. La première étape consiste en une **description systématique du traitement** : nature des données collectées (textes, conversations, métadonnées), sources de données (web scraping, datasets publics, données clients), finalités du traitement (entraînement, fine-tuning, inférence), technologies utilisées (architecture du modèle, framework, infrastructure), acteurs impliqués (développeurs, opérateurs, sous-traitants) et flux de données (collecte, transformation, stockage, transferts). La deuxième étape évalue la **nécessité et la proportionnalité** du traitement : le projet IA est-il réellement nécessaire pour atteindre l'objectif visé ? Les données personnelles pourraient-elles être remplacées par des données synthétiques ou anonymisées ? Le volume de données est-il proportionné aux finalités ? La troisième étape identifie et évalue les **risques pour les droits et libertés** des personnes : risques de ré-identification, de discrimination algorithmique, de décisions automatisées injustes, de fuite de données personnelles via les outputs du modèle, d'utilisation secondaire non autorisée des données.



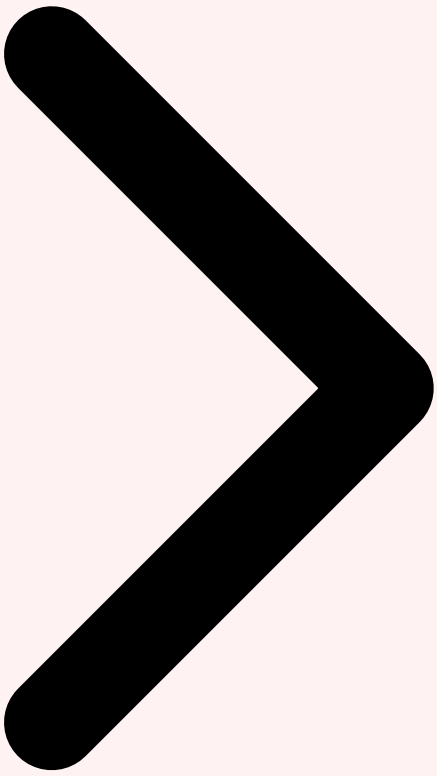
Template DPIA et critères spécifiques IA

Un **template DPIA adapté aux projets d'IA** doit inclure des sections spécifiques qui vont au-delà des modèles classiques. Outre les éléments standard (description du traitement, base légale, mesures de sécurité), le template IA doit couvrir : l'**évaluation des biais algorithmiques** (quels datasets ont été utilisés pour l'entraînement, quels biais potentiels ont été identifiés, quelles mesures de débiaisage ont été mises en œuvre), la **traçabilité du modèle** (version du modèle, provenance des données d'entraînement, model card documentant les capacités et limites), l'**évaluation de la mémorisation** (tests de régurgitation de données personnelles, mesures de la privacy leakage), les **mécanismes de supervision humaine** (qui supervise les outputs, comment les erreurs sont détectées et corrigées, quel est le processus d'escalade), et enfin l'**analyse des risques de sécurité spécifiques à l'IA** (prompt injection, data poisoning, model extraction, evasion attacks). La CNIL met à disposition un logiciel open source — PIA (Privacy Impact Assessment) — que les organisations peuvent utiliser comme base et adapter à leurs besoins spécifiques en matière d'IA.

La **consultation du DPO** (Délégué à la Protection des Données) est requise tout au long du processus de DPIA. Le DPO doit être impliqué dès la phase de conception du projet IA et son avis doit être documenté dans la DPIA. Lorsque la DPIA révèle des risques résiduels élevés que les mesures d'atténuation ne permettent pas de réduire à un niveau acceptable, l'article 36 du RGPD impose une **consultation préalable de la CNIL** avant le démarrage du traitement. Cette consultation est particulièrement pertinente pour les projets IA innovants traitant des données sensibles à grande échelle. En pratique, la CNIL recommande de prendre contact avec ses services le plus en amont possible pour bénéficier d'un accompagnement et éviter les surprises réglementaires en fin de projet. Le délai de réponse de la CNIL à une consultation préalable est de 8 semaines, extensible à 14 semaines pour les dossiers complexes — un facteur à intégrer dans le planning du projet.

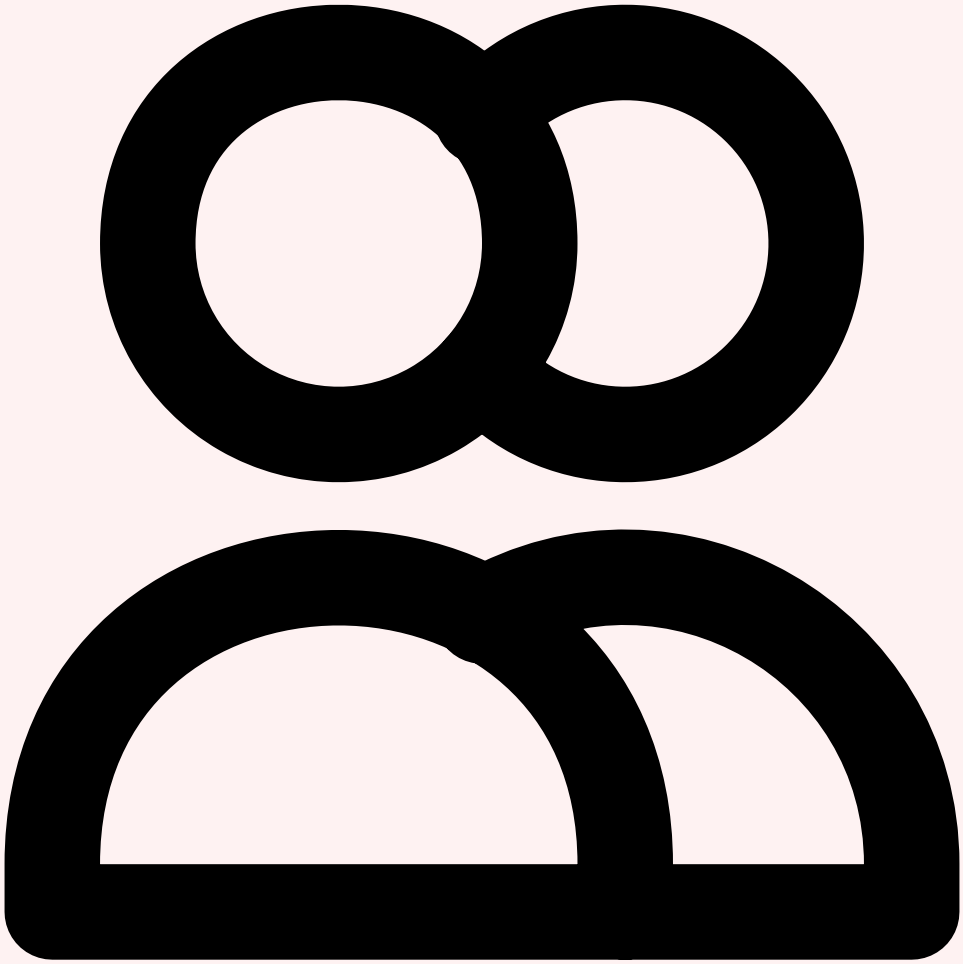


Droit à l'Oubli LLM DPIA Projets IA Décisions Automatisées



6 Décisions Automatisées et Profilage (Article 22)

L'**article 22 du RGPD** établit un principe fondamental : toute personne a le droit de ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé, y compris le profilage, produisant des effets juridiques la concernant ou l'affectant de manière significative. Cette disposition constitue l'un des garde-fous les plus importants face à la montée en puissance de l'IA décisionnelle dans tous les secteurs de l'économie. En 2026, avec la démocratisation des systèmes d'IA capables de prendre des décisions complexes en temps réel — scoring de crédit, tri de candidatures, tarification d'assurance, évaluation des risques de récidive, attribution de prestations sociales —, l'article 22 revêt une pertinence majeur. La question n'est plus de savoir si l'IA peut prendre des décisions affectant les individus, mais dans quelles conditions ces décisions automatisées sont licites et quelles garanties doivent entourer leur mise en oeuvre. Pour approfondir, consultez Responsible Agentic AI : Contrôles, Garde-Fous et Gouvernance.

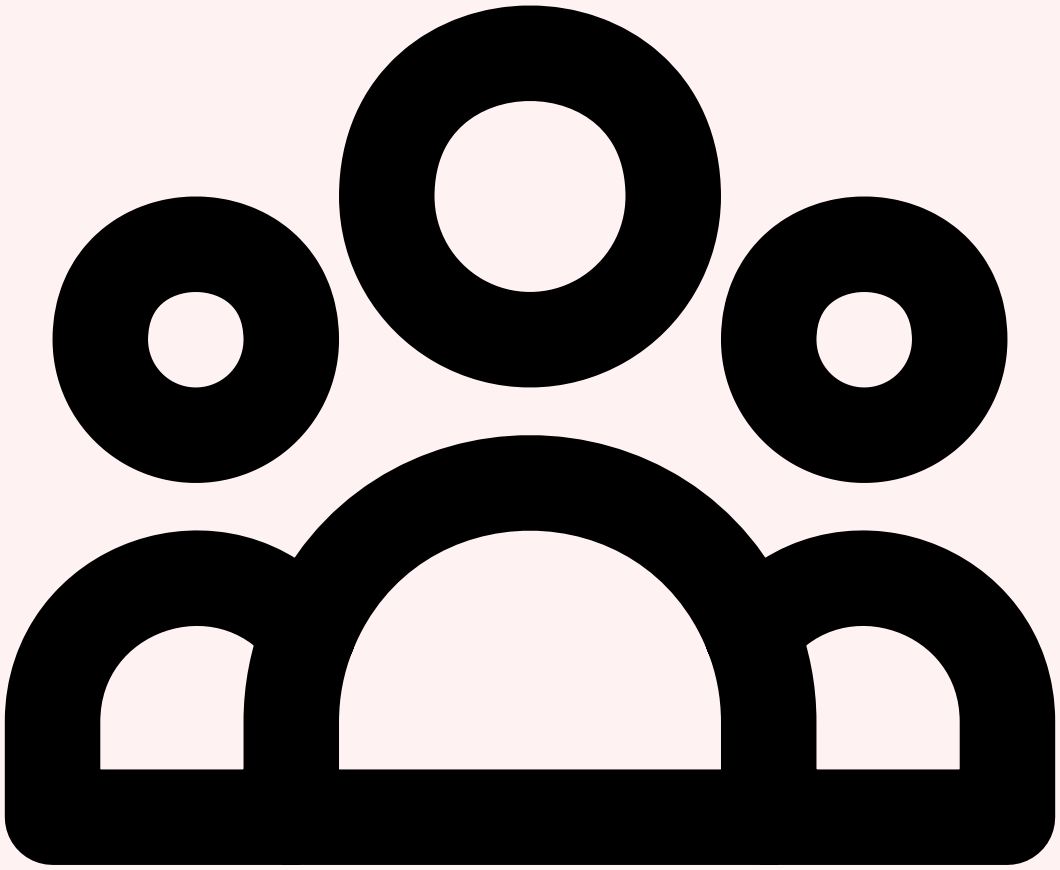


Exceptions et conditions d'application

L'interdiction posée par l'article 22 n'est pas absolue. Le RGPD prévoit **trois exceptions** permettant les décisions automatisées : lorsque la décision est nécessaire à la conclusion ou à l'exécution d'un contrat entre la personne et le responsable du traitement (par exemple, un scoring de crédit automatisé pour l'octroi d'un prêt en ligne), lorsque la décision est autorisée par une législation de l'Union ou d'un État membre (par exemple, la détection automatisée de fraude imposée par la réglementation bancaire), ou lorsque la personne a donné son consentement explicite. Dans tous les cas, le responsable du traitement doit mettre en oeuvre des **mesures de sauvegarde appropriées**, incluant au minimum le droit d'obtenir une intervention humaine, le droit d'exprimer son point de vue et le droit de contester la décision. Ces garanties ne sont pas de simples formalités procédurales : elles doivent être effectives et accessibles, ce qui implique que l'intervention humaine soit réalisée par une personne disposant de l'autorité et de la compétence pour modifier ou annuler la décision algorithmique.

Le droit à l'explication des décisions IA

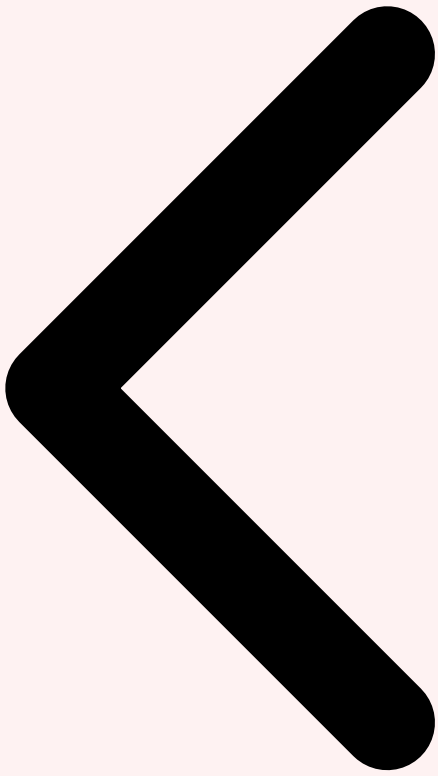
Le **droit à l'explication** constitue un corollaire essentiel de l'article 22 et s'appuie également sur les articles 13, 14 et 15 du RGPD qui imposent de fournir aux personnes concernées des « informations utiles concernant la logique sous-jacente » des traitements automatisés. L'interprétation de cette exigence dans le contexte de l'IA fait l'objet de débats intenses. Pour les systèmes d'IA fondés sur des modèles de boîte noire comme les réseaux de neurones profonds et les LLM, fournir une explication complète de la logique de décision est techniquement difficile. Le domaine de l'**Explainable AI (XAI)** propose des solutions : SHAP (SHapley Additive exPlanations) et LIME (Local Interpretable Model-agnostic Explanations) permettent de décomposer l'influence de chaque variable sur une décision spécifique, les attention maps visualisent les éléments du prompt auxquels le modèle accorde le plus de poids, et les techniques de counterfactual explanations montrent comment la décision aurait changé si certaines variables avaient été différentes. Les autorités de protection des données n'exigent pas une transparence algorithmique totale mais une explication « compréhensible et significative » adaptée au public concerné.



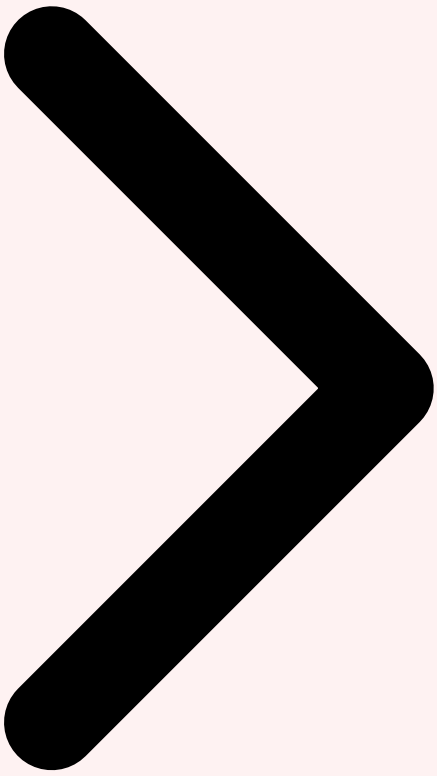
Intervention humaine significative (HITL)

Le concept d'**intervention humaine significative** (Human-in-the-Loop — HITL) est central dans la mise en conformité des systèmes d'IA avec l'article 22. Il ne suffit pas de placer un opérateur humain en bout de chaîne qui se contente de valider mécaniquement les décisions algorithmiques : l'intervention doit être **réelle, éclairée et effective**. L'EDPB a précisé que l'intervention humaine doit être exercée par une personne qui dispose de l'autorité nécessaire pour modifier la décision, qui comprend le fonctionnement du système d'IA et ses limites, et qui effectue réellement un examen individualisé de chaque cas, en prenant en compte les circonstances spécifiques de la personne concernée. Le simple fait qu'un humain clique sur un bouton de validation sans examiner le dossier ne constitue pas une intervention humaine significative au sens du RGPD. Cette exigence a des implications opérationnelles considérables pour les entreprises qui utilisent l'IA pour automatiser des processus décisionnels à haut volume : le dimensionnement des équipes de révision, leur formation aux systèmes d'IA et la mise en œuvre de workflows permettant un examen individualisé dans des délais raisonnables doivent être planifiés dès la conception du système.

Les **cas d'usage à haut risque** illustrent concrètement les enjeux de l'article 22. Dans le **scoring de crédit**, l'utilisation de modèles d'IA pour évaluer la solvabilité des emprunteurs est l'un des domaines les plus régulés, avec des exigences de transparence renforcées par la directive sur le crédit à la consommation et le futur règlement sur l'IA. Le **recrutement assisté par IA** — tri automatisé de CV, analyse vidéo d'entretiens, tests de personnalité algorithmiques — soulève des risques majeurs de discrimination, notamment lorsque les modèles reproduisent les biais historiques présents dans les données d'entraînement. Plusieurs affaires ont mis en lumière des discriminations systémiques fondées sur le genre, l'origine ethnique ou l'âge dans des systèmes de recrutement automatisés. L'**assurance** utilise de plus en plus l'IA pour la tarification personnalisée et l'évaluation des sinistres, avec un risque de discrimination fondée sur des proxies de données sensibles (code postal comme proxy de l'origine ethnique, historique de navigation comme proxy de l'état de santé). Ces cas d'usage requièrent une vigilance particulière et une documentation exhaustive des mesures de sauvegarde mises en œuvre pour protéger les droits des personnes concernées.

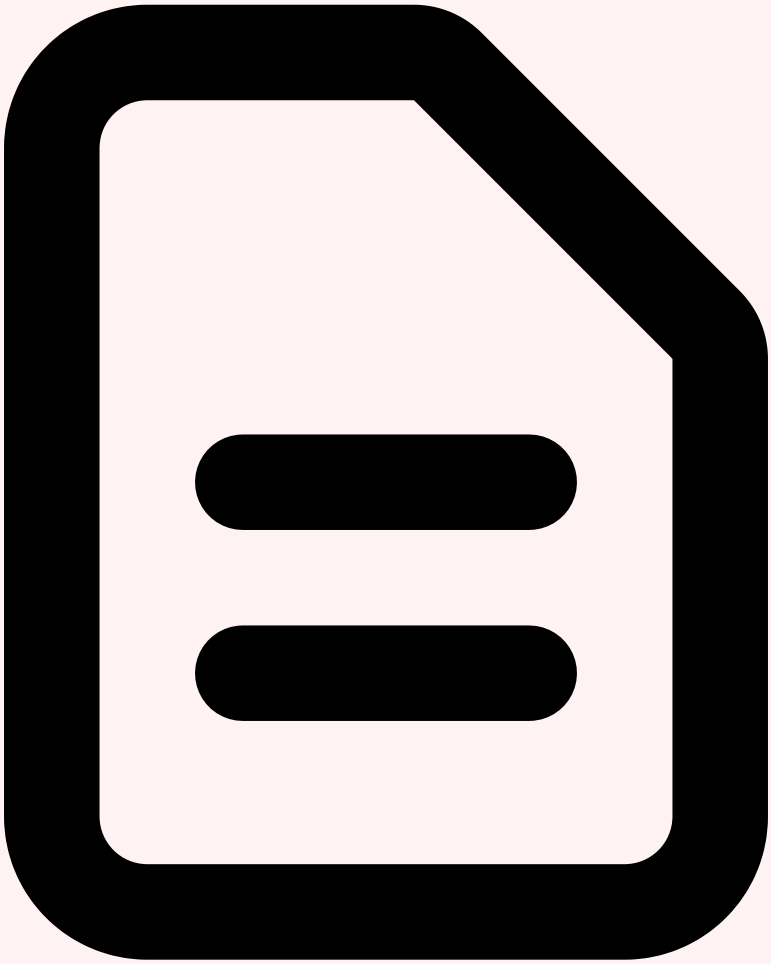


DPIA Projets IA Décisions Automatisées Bonnes Pratiques



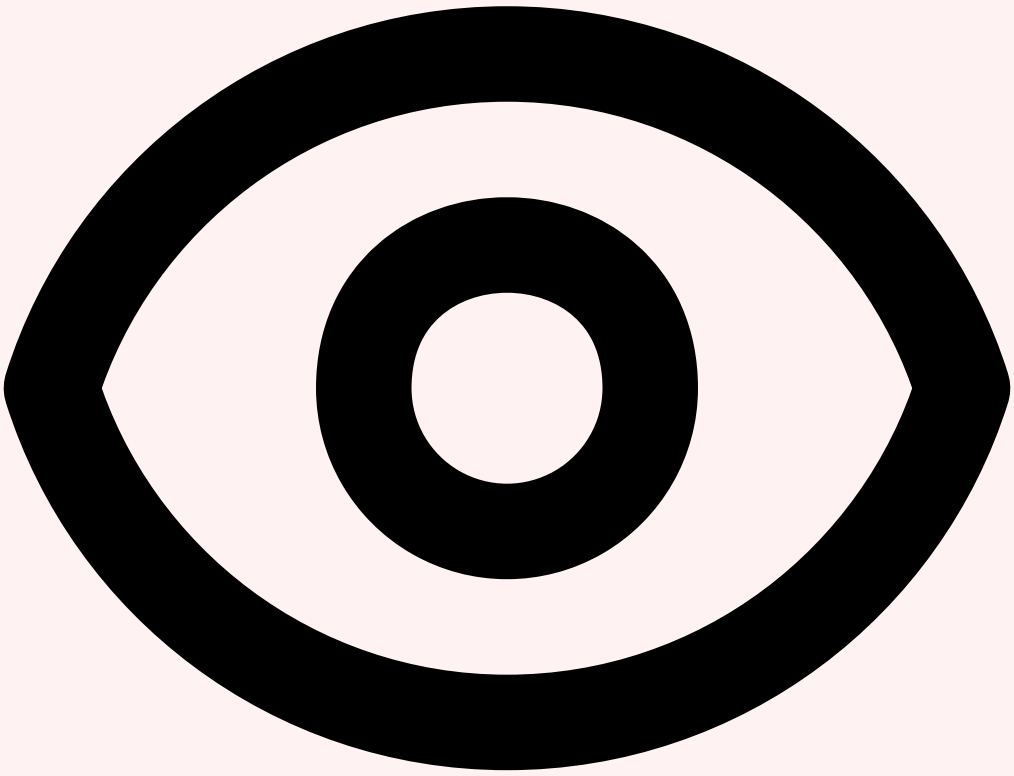
7 Bonnes Pratiques pour la Conformité RGPD/IA

La mise en conformité RGPD des projets d'IA ne se résume pas à un exercice juridique ponctuel : elle nécessite une **approche systémique et continue** qui intègre la protection des données dans la gouvernance, les processus et la culture de l'organisation. En 2026, alors que la convergence entre le RGPD et l'AI Act européen crée un cadre réglementaire de plus en plus exigeant, les entreprises les plus matures en matière de conformité IA ont adopté un ensemble de bonnes pratiques qui vont au-delà du strict respect de la lettre du règlement. Ces bonnes pratiques, issues de l'expérience accumulée par les premiers adoptants de l'IA en entreprise et des retours des autorités de contrôle, constituent un référentiel opérationnel que toute organisation devrait considérer comme le socle minimal de sa démarche de conformité RGPD/IA.



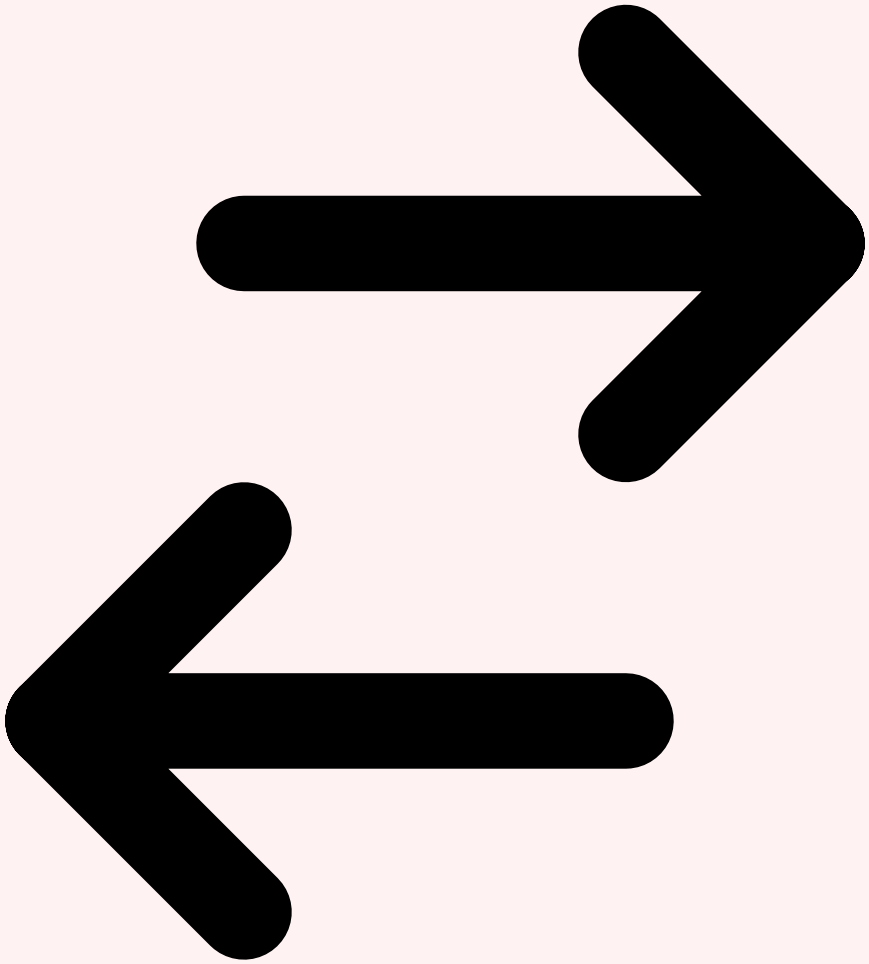
Registre des traitements IA (article 30)

L'**article 30 du RGPD** impose à tout responsable de traitement de tenir un registre des activités de traitement. Pour les traitements impliquant l'IA, ce registre doit être enrichi avec des informations spécifiques : le type de modèle utilisé (modèle propriétaire, open source, fine-tuné), le fournisseur du modèle et les conditions de licence, la provenance des données d'entraînement (datasets publics, données collectées, données synthétiques), les résultats de la DPIA associée, les mesures techniques de protection mises en oeuvre (chiffrement, anonymisation, differential privacy), les guardrails et contrôles d'output implémentés, et les métriques de performance et de biais du modèle. Ce registre enrichi sert de document de référence pour les audits internes et externes, les contrôles de la CNIL, et la démonstration de la conformité dans le cadre de l'accountability (responsabilisation) prévue par l'article 5.2 du RGPD. Les organisations les plus avancées automatisent la mise à jour de ce registre en l'intégrant à leur pipeline MLOps, ce qui garantit que toute modification du modèle ou des données est automatiquement documentée.



Information, transparence et droits des personnes

Les **articles 13 et 14 du RGPD** imposent d'informer les personnes concernées de manière claire et accessible sur le traitement de leurs données. Dans le contexte de l'IA, cette obligation de transparence revêt une dimension particulière car les personnes doivent être informées non seulement que leurs données sont traitées, mais aussi qu'elles le sont par un système automatisé et que des décisions peuvent en découler. Concrètement, la politique de confidentialité doit inclure une section dédiée à l'IA décrivant : les types de traitements IA effectués (entraînement, inférence, personnalisation), les catégories de données utilisées, les finalités spécifiques de chaque traitement IA, l'existence éventuelle de décisions automatisées au sens de l'article 22 et les garanties associées, les droits spécifiques des personnes (opposition à l'entraînement, explication des décisions, intervention humaine), et les coordonnées du DPO pour les réclamations. Les meilleures pratiques incluent également l'utilisation d'**indicateurs visuels** signalant aux utilisateurs quand ils interagissent avec un système d'IA (obligation renforcée par l'AI Act), et la publication de rapports de transparence périodiques détaillant l'utilisation de l'IA par l'organisation.



Sous-traitance IA et transferts internationaux

L'**article 28 du RGPD** encadre strictement la relation entre le responsable de traitement et ses sous-traitants. Lorsqu'une organisation utilise des services d'IA fournis par des tiers — qu'il s'agisse d'API de modèles (OpenAI, Anthropic, Google), de plateformes MLaaS (AWS SageMaker, Azure ML, GCP Vertex AI) ou de solutions SaaS intégrant de l'IA —, un **contrat de sous-traitance conforme à l'article 28** doit être conclu. Ce contrat doit détailler les obligations du sous-traitant en matière de sécurité des données, les instructions de traitement, les conditions de sous-traitance ultérieure, les obligations d'assistance pour répondre aux demandes de droits des personnes, et les conditions de restitution ou de suppression des données en fin de contrat. Les clauses contractuelles doivent spécifiquement aborder la question de l'utilisation des données de prompts pour l'amélioration des modèles du fournisseur — une pratique que de nombreuses entreprises souhaitent interdire contractuellement — et les garanties de confidentialité des données transmises via l'API.

Les **transferts internationaux de données** (articles 44 à 49 du RGPD) constituent un enjeu majeur pour les projets IA, car la majorité des fournisseurs de modèles de fondation sont établis aux États-Unis. Depuis l'invalidation du Privacy Shield par l'arrêt Schrems II en 2020,

les transferts de données vers les États-Unis reposaient sur des clauses contractuelles types (CCT) assorties de mesures supplémentaires. Le **Data Privacy Framework (DPF)**, adopté en 2023, a partiellement résolu cette problématique pour les entreprises américaines certifiées, mais sa pérennité reste incertaine face aux recours juridiques en cours. Les entreprises européennes doivent évaluer pour chaque fournisseur IA : le pays de localisation des serveurs de traitement, l'existence d'une certification DPF ou d'un mécanisme de transfert alternatif, les mesures techniques de protection (chiffrement en transit et au repos, segmentation des données), et la politique de réponse aux demandes d'accès des autorités gouvernementales étrangères. Les options de **déploiement local** — modèles on-premise, instances cloud européennes, modèles open source auto-hébergés — gagnent en popularité comme moyen d'éviter la complexité des transferts internationaux.

Convergence AI Act + RGPD : L'AI Act européen, entré progressivement en application depuis 2024, crée un cadre réglementaire complémentaire au RGPD pour les systèmes d'IA. Les entreprises doivent désormais gérer simultanément la conformité RGPD (protection des données personnelles) et AI Act (sécurité et transparence des systèmes IA). Les synergies sont nombreuses : la DPIA RGPD peut être enrichie pour couvrir les exigences d'évaluation de conformité AI Act, le registre des traitements peut intégrer la classification des risques AI Act, et les mesures de transparence peuvent satisfaire simultanément les deux réglementations. Les organisations qui adoptent une approche intégrée de conformité bénéficient d'économies d'échelle significatives et d'une meilleure cohérence de leur démarche réglementaire. Pour approfondir, consultez IA pour le DFIR : Accélérer les Investigations Forensiques.

La **veille juridique** constitue une activité essentielle dans un paysage réglementaire en constante évolution. Les DPOs et les responsables de conformité doivent suivre non seulement les évolutions législatives (AI Act, propositions de directive sur la responsabilité IA, révisions potentielles du RGPD), mais aussi la jurisprudence émergente, les lignes directrices des autorités de protection des données, les avis de l'EDPB, les décisions de sanction impliquant l'IA, et les normes techniques en cours d'élaboration (ISO/IEC 42001 sur le management de l'IA, ISO/IEC 27701 sur le management de la vie privée). Les organisations doivent appliquer un processus structuré de veille, d'évaluation de l'impact des changements réglementaires sur leurs activités IA, et d'adaptation de leurs pratiques en conséquence. La participation aux consultations publiques, aux groupes de travail sectoriels et aux associations professionnelles constitue également un moyen efficace de rester informé et d'influencer la construction du cadre réglementaire applicable à l'IA.

Ressources open source associées

HF Model RGPD-Expert-1.5B-GGUF HF Dataset rgpd-gdpr-fr HF Space rgpd-gdpr-explorer (démo)

Besoin d'un accompagnement expert ?

Nos consultants en cybersécurité et IA vous accompagnent dans vos projets. Devis personnalisé sous 24h.

Références et ressources externes

- ISO 27001 — Norme internationale de management de la sécurité de l'information
- CNIL — Commission nationale de l'informatique et des libertés
- ENISA — Agence européenne pour la cybersécurité
- OWASP LLM Top 10 — Les 10 risques majeurs pour les applications LLM
- CNIL — Le RGPD — Guide pratique du règlement général sur la protection des données

Pour approfondir ce sujet, consultez notre outil open-source ai-prompt-injection-detector qui facilite la détection des injections de prompt.

Sources et références : ArXiv IA · Hugging Face Papers

FAQ

Qu'est-ce que IA et Conformité RGPD ?

Le concept de IA et Conformité RGPD est détaillé dans les premières sections de cet article, qui couvrent les fondamentaux, les enjeux et le contexte opérationnel. Pour un accompagnement sur ce sujet, contactez nos experts.

Pourquoi IA et Conformité RGPD est-il important en cybersécurité ?

La compréhension de IA et Conformité RGPD permet aux équipes de sécurité d'améliorer leur posture défensive. Les sections « Table des Matières » et « 1 Le RGPD Face au Défi de l'IA Générative » détaillent les raisons de cette importance. Pour un accompagnement sur ce sujet, contactez nos experts.

Comment mettre en œuvre les recommandations de cet article ?

Les recommandations pratiques sont détaillées tout au long de l'article, avec des commandes, des outils et des méthodologies éprouvées. La section « Conclusion » fournit une synthèse actionnable. Pour un accompagnement sur ce sujet, contactez nos experts.

Conclusion

Cet article a couvert les aspects essentiels de Table des Matières, 1 Le RGPD Face au Défi de l'IA Générative, 2 Base Légale du Traitement des Données par l'IA. La mise en pratique de ces recommandations permet de renforcer significativement la posture de securite de votre organisation.

Ayi NEDJIMI Consultants — Expert cybersécurité offensive & intelligence artificielle

ayinedjimi-consultants.fr · ayi@ayinedjimi-consultants.fr