AYI NEDJIMI Consultants

AI Act et LLM : Classifier vos Systèmes IA : Guide Complet

Catégorie : Intelligence Artificielle Lecture : 25 min Publié le : 13/02/2026 Auteur : Ayi NEDJIMI

Guide complet sur l'AI Act européen appliqué aux LLM : classification des systèmes IA par niveau de risque, obligations par catégorie,. Guide.

AI Act et LLM : Classifier vos Systèmes IA : Guide Complet constitue un enjeu majeur pour les professionnels de la sécurité informatique et les équipes techniques. Ce guide détaillé sur ia ai act classifier systemes propose une méthodologie structurée, des outils éprouvés et des recommandations opérationnelles directement applicables. L'objectif est de fournir aux praticiens — consultants, ingénieurs sécurité, administrateurs systèmes — les connaissances et les techniques nécessaires pour aborder ce sujet avec rigueur. Chaque section s'appuie sur des retours d'expérience terrain et intègre les évolutions les plus récentes du domaine. Les recommandations présentées sont adaptées aux environnements d'entreprise et tiennent compte des contraintes opérationnelles réelles.

Table des Matières

Notre avis d'expert

L'IA responsable n'est pas un luxe — c'est une nécessité opérationnelle. Nos audits révèlent que 70% des déploiements IA en entreprise manquent de mécanismes de détection des biais et de garde-fous contre les injections de prompt. Il est temps d'intégrer la sécurité dès la conception des pipelines ML. Guide complet sur l'AI Act européen appliqué aux LLM : classification des systèmes IA par niveau de risque, obligations par catégorie,. Guide. Dans un contexte où l'intelligence artificielle transforme les pratiques de cybersécurité, la maîtrise de ia ai act classifier systemes devient un avantage stratégique pour les équipes techniques. Nous abordons notamment : table des matières, 1 l'ai act : le règlement européen sur l'intelligence artificielle et 2 la pyramide des risques : 4 niveaux de classification. Les professionnels y trouveront des recommandations actionnables, des commandes prêtes à l'emploi et des stratégies de mise en œuvre adaptées aux environnements d'entreprise.

1 L'AI Act : Le Règlement Européen sur l'Intelligence Artificielle

Le **Règlement européen sur l'Intelligence Artificielle** (AI Act), adopté le 13 mars 2024 par le Parlement européen et entré en vigueur le 1er août 2024, constitue le premier cadre juridique complet au monde dédié à la régulation des systèmes d'intelligence artificielle. Ce texte fondateur, composé de 113 articles et 13 annexes, établit un ensemble de règles harmonisées pour le développement, la mise sur le marché et l'utilisation des systèmes IA au sein de l'Union européenne. Pour les organisations déployant des **Large Language Models (LLM)** et des systèmes d'IA générative, la compréhension approfondie de ce règlement est désormais une nécessité stratégique et juridique incontournable.

L'approche retenue par le législateur européen repose sur une **logique de proportionnalité fondée sur le risque**. Contrairement à une interdiction générale ou à une approche sectorielle, l'AI Act classe les systèmes d'IA en quatre niveaux de risque distincts, chacun assorti d'obligations proportionnées. Cette méthodologie s'inspire directement du cadre réglementaire existant pour les produits de sécurité (marquage CE, directives machines) et du RGPD pour la protection des données personnelles. L'objectif est double : protéger les droits fondamentaux des citoyens européens tout en préservant la capacité d'innovation des entreprises et des centres de recherche sur le territoire de l'Union.

Le champ d'application du règlement est particulièrement large. Il concerne les **fournisseurs** (providers) qui développent ou font développer des systèmes d'IA, les **déployeurs** (deployers) qui utilisent ces systèmes dans un contexte professionnel, les **importateurs** et **distributeurs** de solutions IA, ainsi que les fabricants de produits intégrant des composants IA. L'application est extraterritoriale : toute organisation, même établie hors de l'UE, est concernée dès lors que son système IA produit des effets sur le territoire européen. Cette portée rappelle celle du RGPD et oblige les entreprises mondiales à intégrer les exigences européennes dans leur stratégie de conformité globale.

Comment garantir que vos modèles de machine learning ne deviennent pas des vecteurs d'attaque ?

Le calendrier de mise en application est progressif et s'étale jusqu'en 2027. Depuis le **2 février 2025**, les interdictions relatives aux pratiques IA inacceptables sont pleinement applicables. Les obligations concernant les **modèles d'IA à usage général (GPAI)**, catégorie dans laquelle entrent les LLM comme GPT-4, Claude, Gemini ou Mistral, entreront en vigueur le **2 août 2025**. Les exigences relatives aux systèmes à haut risque listés en Annexe III s'appliqueront à partir du **2 août 2026**, tandis que les systèmes à haut risque intégrés dans des produits réglementés (Annexe I) auront jusqu'au **2 août 2027** pour se conformer. Pour les organisations déployant des LLM en production, le compte à rebours a déjà commencé.

Les sanctions prévues par l'AI Act sont significatives et reflètent la volonté du législateur d'assurer l'effectivité du règlement. Les infractions les plus graves, comme l'utilisation de pratiques interdites, peuvent entraîner des amendes allant jusqu'à **35 millions d'euros ou 7% du chiffre d'affaires annuel mondial**. Le non-respect des obligations relatives aux systèmes à haut risque expose à des amendes pouvant atteindre **15 millions d'euros ou 3% du chiffre d'affaires**. Même la fourniture d'informations incorrectes aux autorités peut être sanctionnée à hauteur de **7,5 millions d'euros ou 1% du chiffre d'affaires**. Ces montants, calqués sur le modèle du RGPD, témoignent de l'ambition régulatrice de l'Union européenne en matière d'intelligence artificielle.

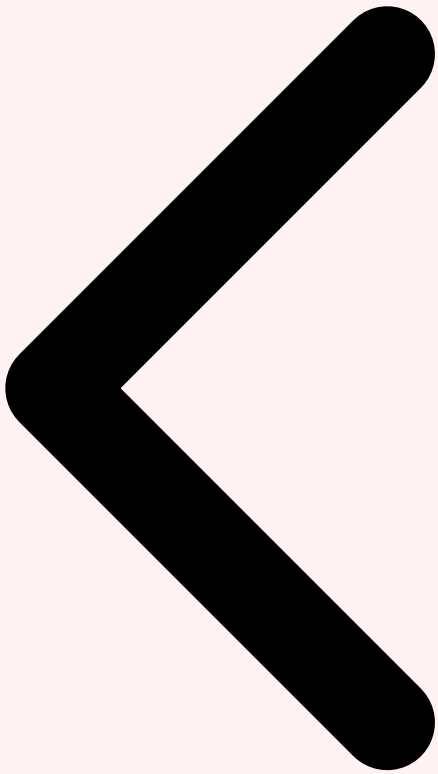
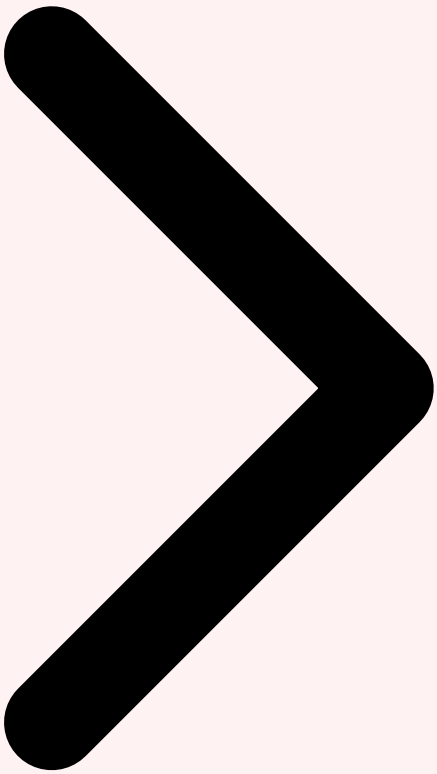


Table des Matières Introduction AI Act Pyramide des Risques



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

Cas concret

En 2023, des chercheurs ont démontré qu'il était possible de manipuler Bing Chat (Copilot) pour exfiltrer des données personnelles via des techniques d'injection de prompt indirecte. Cette attaque exploitait la capacité du LLM à accéder aux résultats de recherche web, transformant un assistant en vecteur d'exfiltration.

2 La Pyramide des Risques : 4 Niveaux de Classification

Le cœur de l'AI Act repose sur une **classification pyramidale à quatre niveaux de risque** qui détermine l'intensité des obligations réglementaires applicables à chaque système d'IA. Cette approche graduée constitue l'innovation juridique majeure du règlement : elle évite l'écueil d'une réglementation uniforme et inadaptée en modulant les exigences en fonction de l'impact potentiel du système sur les droits fondamentaux et la sécurité des personnes. Pour les équipes techniques déployant des LLM, comprendre précisément où se situe chaque cas d'usage dans cette pyramide est la première étape indispensable de toute démarche de conformité.

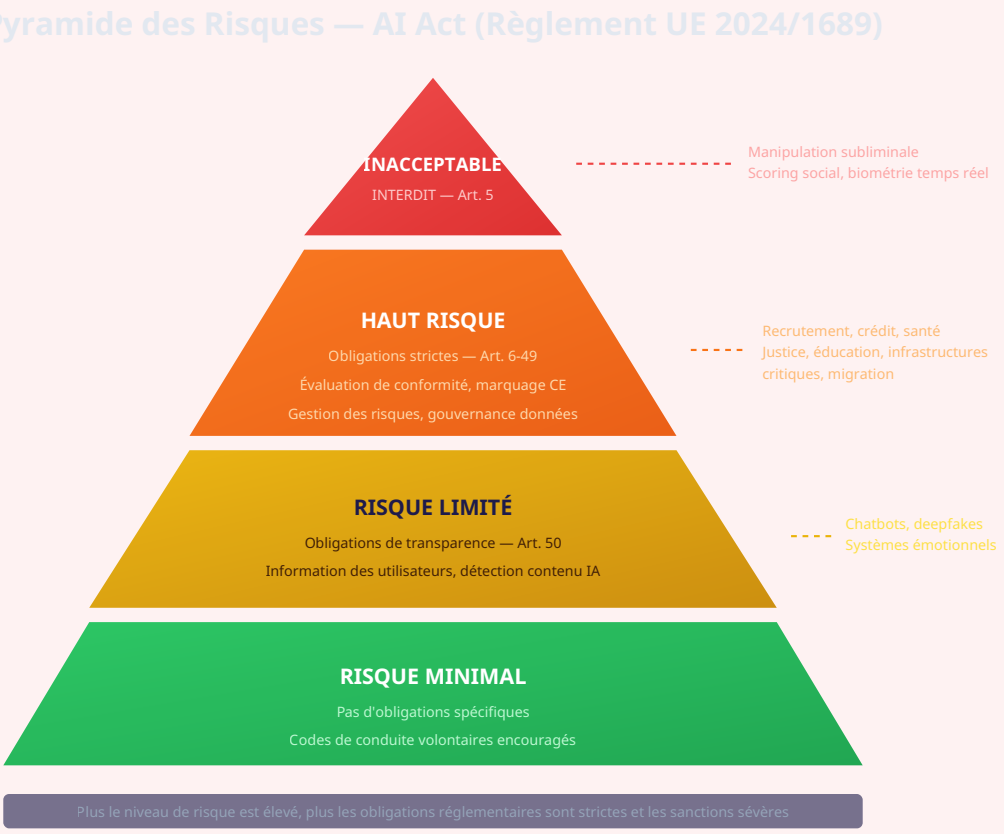


Figure 1 — Pyramide des 4 niveaux de risque définis par l'AI Act (Règlement UE 2024/1689)

Niveau 1 : Risque Inacceptable — Les Pratiques Interdites (Article 5)

Au sommet de la pyramide se trouvent les **pratiques IA strictement interdites** par l'article 5 du règlement, applicables depuis le 2 février 2025. Ces interdictions couvrent huit catégories de systèmes considérés comme portant atteinte de manière intolérable aux droits fondamentaux. Parmi elles, on trouve les systèmes utilisant des **techniques subliminales ou manipulatrices** pour altérer le comportement d'une personne de manière à causer un préjudice significatif, les systèmes exploitant les vulnérabilités liées à l'âge, au handicap ou à la situation sociale, ainsi que les systèmes de **notation sociale**

(social scoring) par les autorités publiques. L'identification biométrique en temps réel dans l'espace public est également interdite, sauf exceptions très encadrées pour les forces de l'ordre. Pour les LLM, cela signifie concrètement que tout système conçu pour manipuler psychologiquement les utilisateurs ou exploiter leurs vulnérabilités est hors-la-loi, quelle que soit la sophistication technique employée. Pour approfondir, consultez Vecteurs en Intelligence Artificielle.

Niveau 2 : Haut Risque — Le Cœur du Dispositif (Articles 6 à 49)

La catégorie **haut risque** représente le cœur opérationnel du règlement et concentre l'essentiel des obligations de conformité. Un système d'IA est classé à haut risque dans deux cas : soit il est intégré comme composant de sécurité dans un produit déjà couvert par la législation harmonisée de l'UE (Annexe I : dispositifs médicaux, machines, jouets, équipements radio, aviation civile), soit il entre dans l'une des huit catégories sensibles listées en **Annexe III**. Ces catégories incluent l'identification biométrique et la catégorisation des personnes physiques, la gestion et l'exploitation des infrastructures critiques (énergie, transports, eau, télécommunications), l'éducation et la formation professionnelle (accès, évaluation, orientation), l'emploi et la gestion des travailleurs (recrutement, promotion, licenciement), l'accès aux services publics essentiels et aux prestations sociales, les activités répressives et judiciaires, la gestion des migrations et du contrôle aux frontières, ainsi que l'administration de la justice et les processus démocratiques.

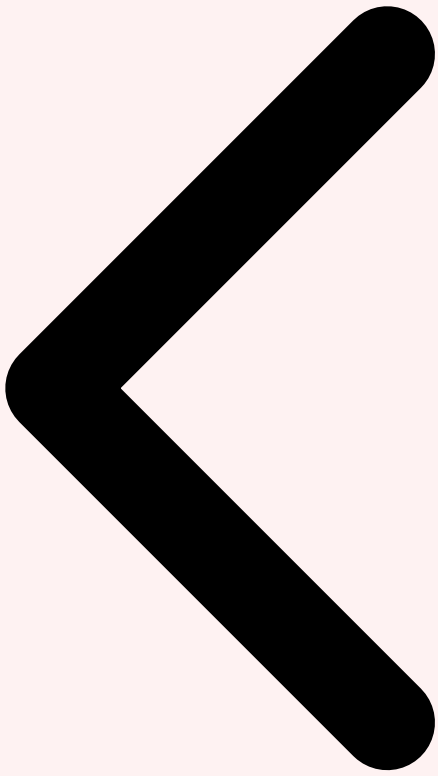
Avez-vous évalué les risques d'injection de prompt sur vos systèmes d'IA en production ?

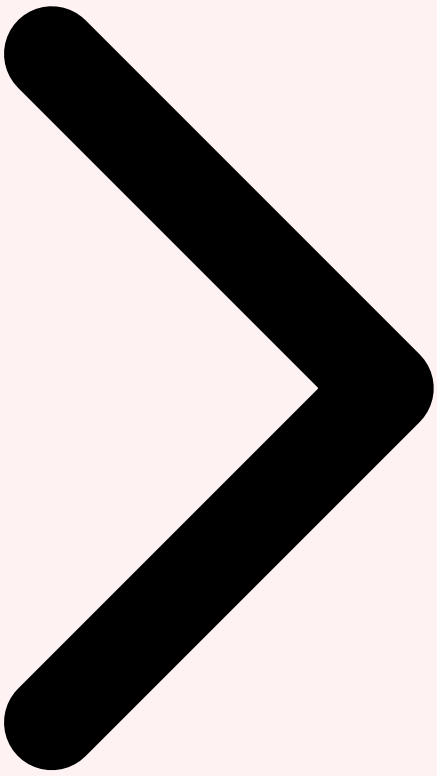
Niveau 3 : Risque Limité — L'Obligation de Transparence (Article 50)

Le troisième niveau concerne les systèmes présentant un **risque limité**, pour lesquels l'obligation principale est celle de transparence définie à l'article 50. Les systèmes IA conçus pour interagir directement avec des personnes physiques, comme les **chatbots et assistants conversationnels**, doivent clairement informer l'utilisateur qu'il communique avec un système d'IA, sauf si cela est évident au vu des circonstances. Les systèmes générant des contenus synthétiques (texte, image, audio, vidéo) doivent marquer ces contenus de manière lisible par machine, conformément aux normes techniques qui seront définies par les organismes de normalisation européens. Les systèmes de reconnaissance des émotions et de catégorisation biométrique doivent informer les personnes exposées. Cette catégorie est particulièrement pertinente pour les LLM déployés en interface utilisateur : tout chatbot alimenté par GPT-4, Claude, Gemini ou un modèle open source doit signaler sa nature artificielle.

Niveau 4 : Risque Minimal — Liberté Encadrée

La base de la pyramide, la plus large, couvre les systèmes IA à **risque minimal ou nul**, qui représentent la grande majorité des applications IA actuellement déployées. Les filtres anti-spam, les systèmes de recommandation de contenu, les outils d'optimisation logistique ou les moteurs de recherche augmentés par l'IA entrent typiquement dans cette catégorie. Aucune obligation spécifique n'est imposée par le règlement pour ces systèmes, mais le législateur **encourage l'adoption volontaire de codes de conduite** reprenant les bonnes pratiques en matière de transparence, d'équité et de robustesse. Pour les organisations, classer un système dans cette catégorie ne dispense pas pour autant de respecter les autres réglementations applicables, notamment le RGPD pour le traitement des données personnelles ou les directives sectorielles spécifiques.





3 Classifier vos Systèmes LLM

La classification d'un système basé sur un LLM sous l'AI Act ne dépend pas du modèle sous-jacent lui-même, mais de **l'usage spécifique qui en est fait** et du contexte de déploiement. C'est une distinction fondamentale que beaucoup d'organisations peinent encore à appréhender : un même modèle GPT-4 ou Claude peut être à risque minimal lorsqu'il est utilisé comme assistant de rédaction interne, à risque limité lorsqu'il alimente un chatbot client, et à haut risque lorsqu'il participe à un processus de décision en matière de recrutement ou d'évaluation de crédit. La classification se fait donc au niveau du **système d'IA** (l'application complète) et non au niveau du modèle de fondation.

Méthodologie de Classification en 5 Étapes

Pour classifier correctement un système LLM, nous recommandons une approche structurée en cinq étapes. **Premièrement**, identifiez précisément la finalité du système : quel est l'objectif métier, quelle décision ou action le système influence-t-il, et qui sont les personnes affectées par son fonctionnement ? **Deuxièmement**, vérifiez si le cas d'usage

tombe sous le coup des pratiques interdites de l'article 5 — si oui, le projet doit être abandonné ou fondamentalement repensé. **Troisièmement**, examinez si le système est intégré comme composant de sécurité d'un produit couvert par l'Annexe I ou s'il entre dans l'une des catégories sensibles de l'Annexe III. **Quatrièmement**, évaluez si le système interagit directement avec des utilisateurs ou génère du contenu synthétique, ce qui le placerait au niveau de risque limité. **Cinquièmement**, documentez votre raisonnement de classification de manière traçable et auditable, car les autorités de surveillance pourront contester une classification jugée inadéquate.

Cas d'Usage Typiques et Leur Classification

Examinons les cas d'usage les plus courants des LLM en entreprise et leur classification probable. Les **assistants de rédaction et de synthèse documentaire** internes, sans interaction directe avec des tiers et sans impact décisionnel, relèvent généralement du risque minimal. Les **chatbots de service client** alimentés par un LLM sont au minimum à risque limité en raison de l'obligation de transparence, et potentiellement à haut risque s'ils prennent des décisions ayant un impact juridique sur les consommateurs (refus de remboursement, résiliation automatique). Les systèmes de **screening automatisé de CV** ou de pré-sélection de candidats sont systématiquement à haut risque (Annexe III, point 4). Les outils de **scoring de crédit** ou d'évaluation de risque d'assurance intégrant un LLM sont à haut risque (Annexe III, point 5b). Les **systèmes d'aide au diagnostic médical** exploitant un LLM pour interpréter des données cliniques sont à haut risque au double titre de l'Annexe I (dispositifs médicaux) et de l'Annexe III.

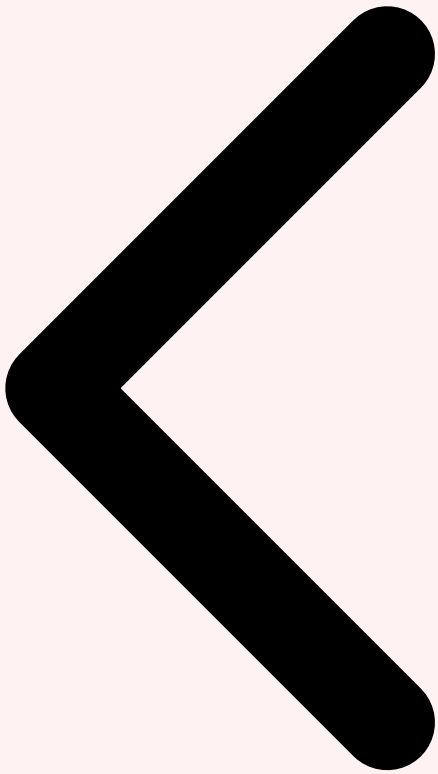
Le Piège de la Clause d'Exception (Article 6, paragraphe 3)

L'article 6, paragraphe 3, de l'AI Act introduit une **clause d'exception** pour les systèmes listés en Annexe III qui ne posent pas de risque significatif de préjudice pour la santé, la sécurité ou les droits fondamentaux des personnes physiques. Un système peut échapper à la classification à haut risque s'il remplit l'une des conditions suivantes : il effectue une tâche procédurale étroite, il améliore le résultat d'une activité humaine préalable, il détecte des schémas décisionnels sans remplacer ni influencer l'évaluation humaine, ou il effectue une tâche préparatoire à une évaluation pertinente pour les cas d'utilisation de l'Annexe III. Cependant, cette exception ne s'applique **jamais** aux systèmes effectuant du profilage de personnes physiques. Les organisations tentées d'utiliser cette clause pour éviter les obligations de haut risque doivent être extrêmement prudentes : elles doivent documenter leur raisonnement et le notifier à l'autorité compétente **avant** la mise sur le marché du système. Toute contestation par l'autorité de surveillance peut entraîner une reclassification et l'application rétroactive des obligations.

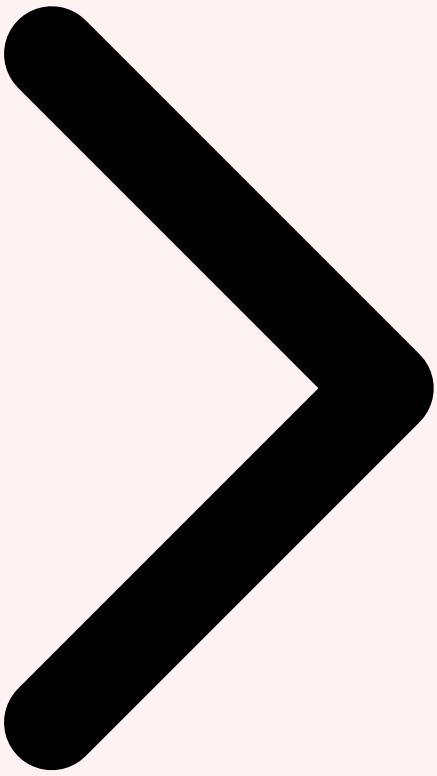
Double Classification : LLM Multi-usages

Un défi spécifique aux LLM réside dans la **multiplicité des cas d'usage** d'une même infrastructure. Une plateforme d'IA d'entreprise construite sur un modèle unique peut servir simultanément à la synthèse documentaire (risque minimal), au support client (risque limité) et à l'aide au recrutement (haut risque). Dans ce scénario, chaque cas

d'usage doit être évalué individuellement. La recommandation pratique est de **segmenter les déploiements** par niveau de risque, avec des contrôles et une documentation proportionnés à chaque usage. Cette segmentation facilite non seulement la conformité, mais permet également d'optimiser les investissements en gouvernance en concentrant les efforts sur les systèmes à haut risque tout en maintenant une approche légère pour les usages à risque minimal. Pour approfondir, consultez Détection Proactive de Contenu Généré par IA Multimodal.



Pyramide des Risques Classifier vos LLM GPAI Modèles Fondation



4 GPAI : Obligations pour les Modèles de Fondation

L'une des innovations majeures de l'AI Act est l'introduction d'un cadre spécifique pour les **modèles d'IA à usage général (General-Purpose AI ou GPAI)**, définis aux articles 51 à 56 du règlement. Cette catégorie vise directement les modèles de fondation tels que GPT-4, Claude, Gemini, Llama, Mistral ou tout autre LLM capable d'accomplir une grande variété de tâches distinctes, indépendamment de la manière dont il est mis sur le marché. La définition retenue par le règlement est fonctionnelle : un modèle GPAI est un modèle d'IA qui présente une **généralité significative**, qui est capable d'exécuter de manière compétente un large éventail de tâches distinctes, et qui peut être intégré dans une variété de systèmes ou d'applications en aval. Cette définition englobe de facto tous les grands modèles de langage actuels.

Obligations de Base pour Tous les Modèles GPAI (Article 53)

Tout fournisseur de modèle GPAI, qu'il soit open source ou propriétaire, doit respecter un ensemble d'**obligations minimales** définies à l'article 53. Premièrement, il doit établir et maintenir à jour une **documentation technique** du modèle et de son processus d'entraînement, qu'il met à disposition de l'AI Office et des autorités nationales compétentes sur demande. Cette documentation doit être suffisamment détaillée pour permettre aux fournisseurs en aval d'exercer leurs propres obligations de conformité. Deuxièmement, il doit mettre en place une **politique de respect du droit d'auteur** de l'Union européenne, incluant l'identification et le respect des réservations de droits formulées par les titulaires de droits au titre de la directive (UE) 2019/790 sur le droit d'auteur. Troisièmement, il doit publier un **résumé suffisamment détaillé des données d'entraînement** utilisées pour le modèle, selon un modèle fourni par l'AI Office. Cette dernière obligation est particulièrement sensible pour les développeurs de LLM, car elle touche à la transparence sur les corpus d'entraînement, sujet de contentieux actifs entre les éditeurs de presse et les développeurs de modèles.

Modèles GPAI à Risque Systémique (Article 51, paragraphe 2)

L'AI Act crée une sous-catégorie de modèles GPAI présentant un **risque systémique**, soumis à des obligations renforcées. Un modèle est présumé à risque systémique lorsque la quantité cumulée de calcul utilisée pour son entraînement dépasse **10^25 FLOP** (floating point operations), soit environ 10 milliards de milliards de milliards d'opérations. Ce seuil quantitatif, bien qu'imparfait, vise à identifier les modèles les plus puissants et potentiellement les plus dangereux. La Commission européenne peut également désigner un modèle comme présentant un risque systémique sur la base d'autres critères qualitatifs, tels que le nombre d'utilisateurs professionnels enregistrés, le nombre de paramètres du modèle, la taille et les caractéristiques des données d'entraînement, ou les capacités du modèle en termes de génération de contenu, de résolution de problèmes ou d'interaction multimodale. En pratique, les modèles GPT-4, Gemini Ultra et potentiellement Claude Opus dépassent ce seuil de 10^25 FLOP et seraient donc qualifiés de GPAI à risque systémique.

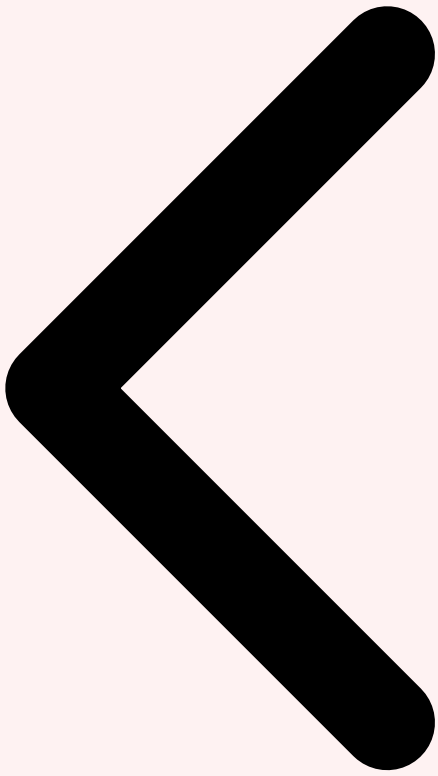
Obligations Renforcées pour les GPAI à Risque Systémique (Article 55)

Les fournisseurs de modèles GPAI à risque systémique doivent, en plus des obligations de base, satisfaire à des **exigences supplémentaires significatives**. Ils doivent réaliser des **évaluations de modèle** (model evaluations) conformément à des protocoles standardisés, incluant la conduite et la documentation de tests adversariaux (red teaming) du modèle en vue d'identifier et d'atténuer les risques systémiques. Ils doivent évaluer et atténuer les risques systémiques possibles, y compris leurs sources, au niveau de l'Union. Ils doivent **suivre, documenter et signaler** sans retard injustifié à l'AI Office et, le cas échéant, aux autorités nationales compétentes, les incidents graves et les mesures correctives possibles pour y remédier. Ils doivent également assurer un **niveau adéquat de protection en**

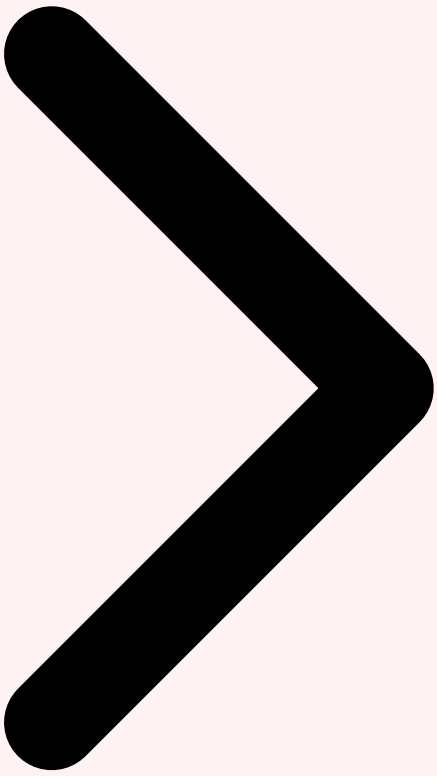
matière de cybersécurité pour le modèle et son infrastructure physique. Ces obligations placent les développeurs des plus grands LLM dans un régime de surveillance renforcé comparable à celui des institutions financières systémiques.

L'Exception Open Source et Ses Limites

Le règlement accorde un **traitement allégé aux modèles GPAI open source** (article 53, paragraphe 2), reconnaissant leur contribution à l'innovation et à la recherche. Les fournisseurs de modèles GPAI dont les paramètres, y compris les poids, l'architecture du modèle et les informations relatives à l'utilisation du modèle, sont rendus publiquement accessibles sous une licence libre et open source, sont exemptés de la plupart des obligations de l'article 53, à l'exception de l'obligation de publier le résumé des données d'entraînement et de la politique de droit d'auteur. Cependant, cette exception **ne s'applique pas aux modèles à risque systémique** : un modèle open source dépassant le seuil de 10^25 FLOP reste soumis à l'intégralité des obligations renforcées. Cette nuance est importante pour des modèles comme Llama (Meta) ou les futurs modèles open source de grande taille, qui ne pourront pas s'abriter derrière leur licence libre pour échapper à la surveillance réglementaire. La définition même d'open source dans le contexte de l'IA reste débattue : publier les poids d'un modèle sans les données d'entraînement ni le code d'entraînement constitue-t-il réellement de l'open source au sens de l'AI Act ? L'AI Office devra clarifier cette question dans ses futures lignes directrices.



Classifier vos LLM GPAI Modèles Fondation Obligations Haut Risque



5 Obligations pour les Systèmes à Haut Risque

Les systèmes d'IA classés à **haut risque** sont soumis aux obligations les plus exigeantes du règlement, détaillées aux articles 8 à 49 de l'AI Act. Ces obligations s'appliquent principalement aux fournisseurs (developers), mais certaines incombent également aux déployeurs (users) et aux importateurs/distributeurs. Pour les organisations déployant des LLM dans des contextes à haut risque — recrutement, crédit, santé, éducation, justice — la mise en conformité requiert une transformation profonde des processus de développement, de test et de gouvernance des systèmes d'IA.

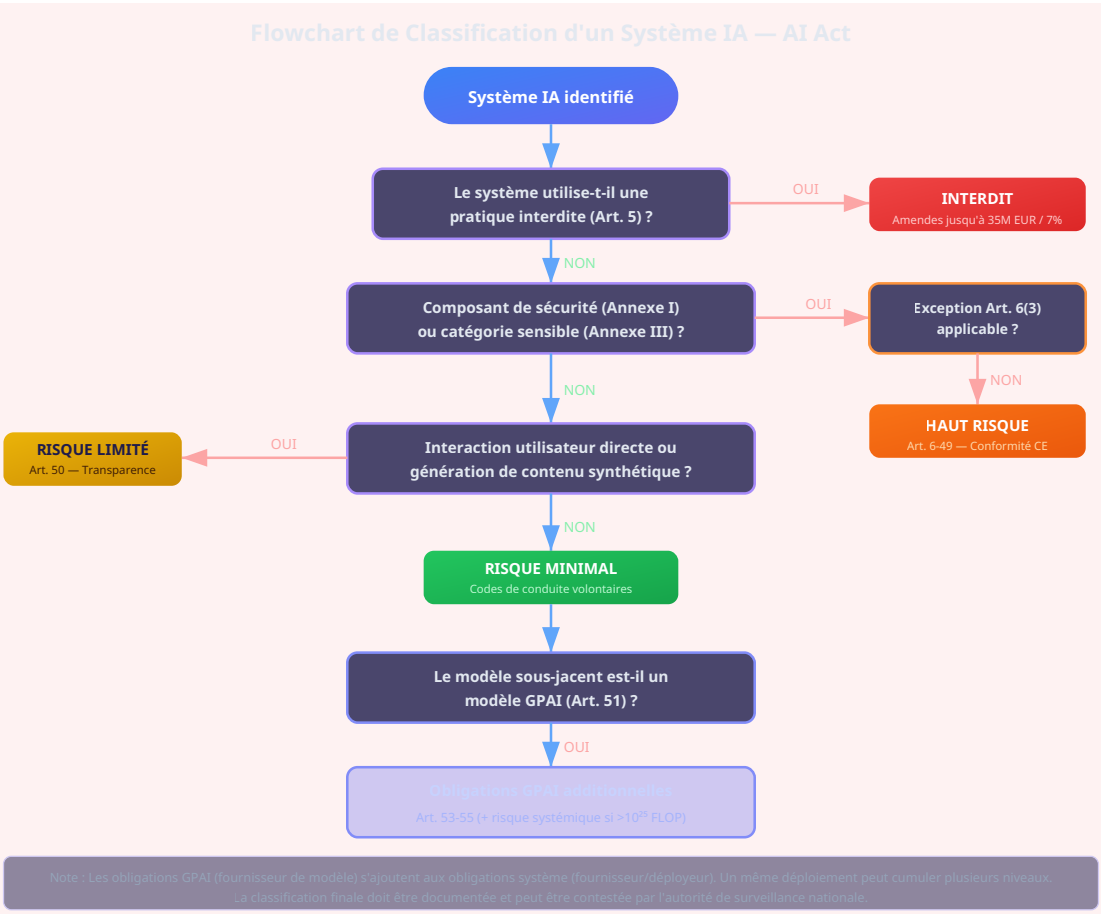


Figure 2 — Flowchart de classification d'un système IA selon les articles 5, 6 et 50 de l'AI Act

Système de Gestion des Risques (Article 9)

L'article 9 impose la mise en œuvre d'un **système de gestion des risques** continu et itératif, couvrant l'intégralité du cycle de vie du système IA. Ce système doit identifier et analyser les risques connus et raisonnablement prévisibles que le système peut poser pour la santé, la sécurité ou les droits fondamentaux. Il doit estimer et évaluer ces risques dans des conditions d'utilisation normale et de mauvaise utilisation raisonnablement prévisible. Il doit définir des mesures de gestion des risques appropriées, puis évaluer l'efficacité de ces mesures. Pour un LLM de recrutement, cela signifie concrètement documenter les risques de biais discriminatoires (genre, origine ethnique, âge, handicap), les risques d'hallucination du modèle pouvant fausser l'évaluation d'un candidat, les risques d'attaque adversariale visant à manipuler les résultats, et les mesures d'atténuation mises en œuvre pour chacun de ces risques. Ce processus doit être documenté, mis à jour régulièrement et intégré dans le système de management de la qualité de l'organisation. Pour approfondir, consultez IA pour l'Analyse de Logs et Détection d'Anomalies en Temps Réel.

Gouvernance des Données (Article 10)

L'article 10 établit des exigences strictes en matière de **gouvernance des données** utilisées pour l'entraînement, la validation et le test des systèmes IA à haut risque. Les jeux de données doivent être pertinents, suffisamment représentatifs, exempts d'erreurs dans la mesure du possible, et complets au regard de la finalité prévue du système. Les données

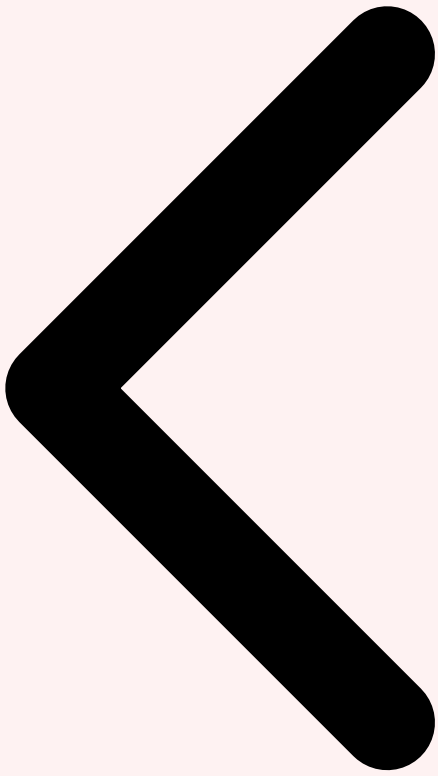
d'entraînement doivent tenir compte des caractéristiques géographiques, contextuelles, comportementales et fonctionnelles spécifiques au cadre dans lequel le système est destiné à être utilisé. Des pratiques appropriées de **détection et correction des biais** doivent être mises en œuvre. Pour les LLM, cette obligation est particulièrement complexe car les modèles de fondation sont généralement entraînés sur des corpus massifs (des centaines de téraoctets de texte web) dont le contenu exact est difficile à auditer. Les déployeurs utilisant un modèle GPAI tiers devront s'appuyer sur la documentation technique fournie par le fournisseur du modèle et compléter par leurs propres évaluations sur les données spécifiques à leur cas d'usage.

Documentation Technique et Traçabilité (Articles 11-12)

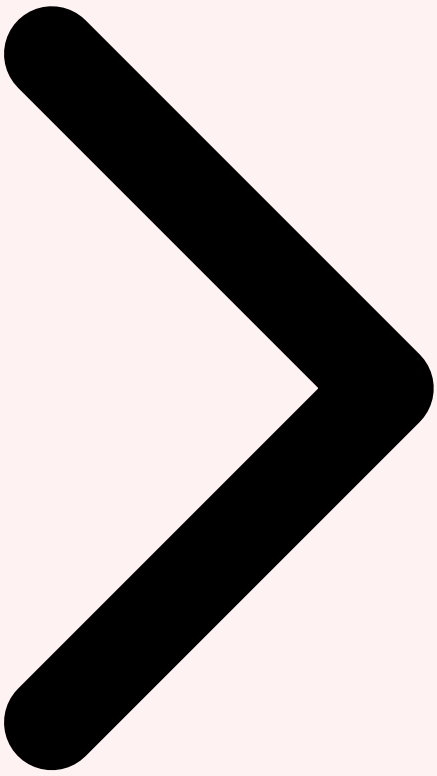
Les articles 11 et 12 imposent la tenue d'une **documentation technique exhaustive** et la mise en œuvre de mécanismes de **journalisation automatique (logging)**. La documentation technique, dont le contenu minimum est précisé en Annexe IV, doit inclure une description générale du système IA, une description détaillée de ses éléments et de son processus de développement, des informations sur les données d'entraînement et de test, les métriques de performance utilisées, une description du système de gestion des risques, et les modifications apportées tout au long du cycle de vie. Les journaux (logs) doivent permettre le traçage des opérations du système pendant toute sa période d'utilisation, avec une granularité suffisante pour identifier les situations de risque et faciliter la surveillance post-commercialisation. Pour un LLM, cela implique de logger chaque requête et chaque réponse, les métadonnées contextuelles, les scores de confiance et les éventuelles interventions de filtrage, conformément aux exigences du RGPD en matière de minimisation des données.

Surveillance Humaine (Article 14)

L'article 14 constitue l'une des dispositions les plus structurantes pour les déploiements de LLM : l'obligation de **surveillance humaine (human oversight)**. Les systèmes IA à haut risque doivent être conçus de telle sorte qu'ils puissent être surveillés efficacement par des personnes physiques pendant leur période d'utilisation. Les mesures de surveillance humaine doivent permettre à la personne en charge de comprendre correctement les capacités et les limites du système, de détecter et corriger les anomalies, les dysfonctionnements et les performances inattendues, de pouvoir décider de ne pas utiliser le système ou d'interrompre, annuler ou inverser le résultat produit par le système. Concrètement, pour un LLM utilisé dans un processus de recrutement, cela signifie qu'un être humain qualifié doit toujours pouvoir **examiner, contester et renverser** toute recommandation produite par le système avant qu'elle ne produise des effets juridiques sur le candidat. Le concept de "human-in-the-loop" ne suffit pas : l'AI Act exige un contrôle humain effectif et significatif, pas une simple validation automatique.



GPAI Modèles Fondation Obligations Haut Risque Conformité Pratique



6 Conformité en Pratique : Documentation et Audit

La mise en conformité avec l'AI Act ne se résume pas à une classification théorique des systèmes : elle exige la production d'une **documentation technique structurée**, la mise en œuvre de processus d'audit vérifiables et l'intégration de mécanismes de gouvernance dans les workflows existants de l'organisation. Pour les équipes déployant des LLM, cette dimension pratique est souvent sous-estimée et constitue pourtant le défi opérationnel majeur de la conformité. L'expérience du RGPD a montré que la documentation et la démonstration de conformité (accountability) représentent une charge de travail significative, et l'AI Act reprend cette logique en l'appliquant spécifiquement aux systèmes d'intelligence artificielle.

Le Dossier Technique (Annexe IV)

L'Annexe IV de l'AI Act détaille le contenu minimum de la **documentation technique** requise pour les systèmes à haut risque. Cette documentation doit être préparée avant la mise sur le marché ou la mise en service du système et doit être maintenue à jour tout au

long de son cycle de vie. Elle comprend une description générale du système IA incluant sa finalité prévue, le nom du fournisseur, la version du système et les interactions avec d'autres systèmes. Elle doit contenir une description détaillée des **éléments du système** : l'architecture du modèle, les algorithmes utilisés, les choix de conception clés, les hypothèses formulées et les compromis réalisés. La documentation des **données d'entraînement, de validation et de test** est requise : description des jeux de données, origine des données, méthodes de collecte, étiquetage, nettoyage et enrichissement, taille des jeux de données, caractéristiques pertinentes et lacunes connues. Pour un LLM basé sur un modèle GPAI tiers, le déployeur peut s'appuyer sur la documentation fournie par le fournisseur du modèle, mais doit compléter avec la documentation spécifique à son fine-tuning, son RAG (Retrieval-Augmented Generation) et son contexte applicatif propre.

Évaluation de Conformité (Articles 43-44)

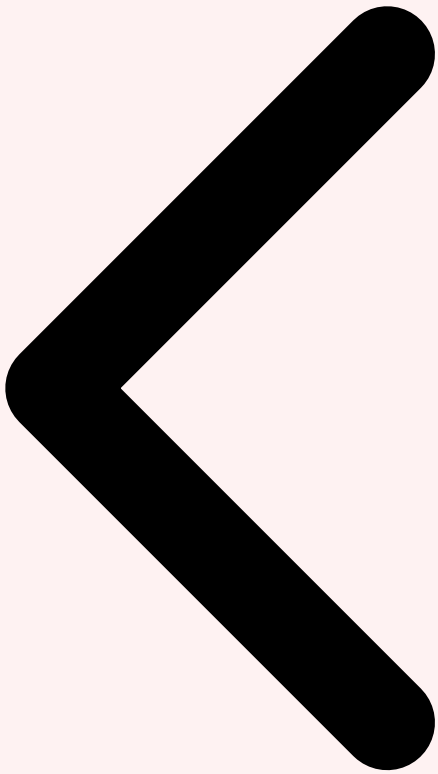
Avant la mise sur le marché d'un système IA à haut risque, le fournisseur doit procéder à une **évaluation de conformité** selon l'une des procédures prévues aux articles 43 et 44. Pour la majorité des systèmes à haut risque relevant de l'Annexe III, l'évaluation de conformité peut être réalisée par le fournisseur lui-même (auto-évaluation) selon la procédure de contrôle interne décrite à l'Annexe VI. Cette procédure exige la vérification de la conformité du système de management de la qualité aux exigences de l'article 17, l'examen de la documentation technique pour démontrer la conformité aux exigences des articles 8 à 15, et la vérification que le système est conforme aux spécifications qui lui sont applicables. Cependant, pour les systèmes IA utilisés dans le cadre de l'**identification biométrique à distance**, l'évaluation doit être réalisée par un **organisme notifié** (organisme tiers accrédité), ce qui implique des coûts et des délais supplémentaires significatifs. Une fois l'évaluation de conformité réussie, le fournisseur appose le **marquage CE** sur le système et établit une déclaration UE de conformité.

Système de Management de la Qualité (Article 17)

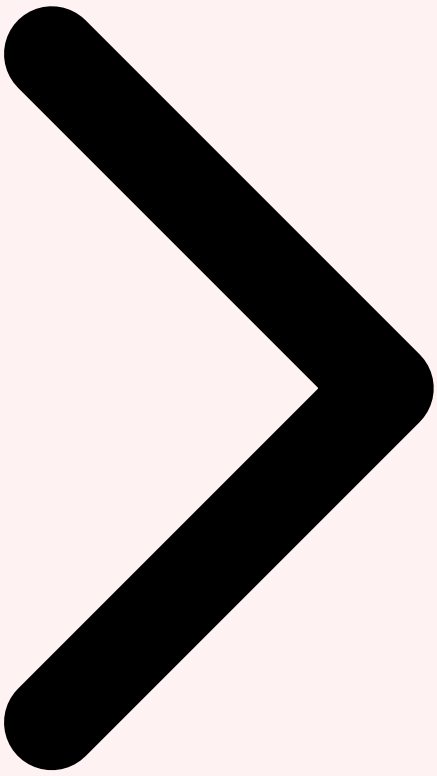
L'article 17 impose aux fournisseurs de systèmes IA à haut risque la mise en œuvre d'un **système de management de la qualité (SMQ)** documenté et systématique. Ce SMQ doit couvrir la stratégie de conformité réglementaire, les techniques et procédures de conception et de développement, les procédures de test et de validation (y compris avant et après déploiement), les spécifications techniques et les normes utilisées, les systèmes et procédures de gestion des données, le système de gestion des risques, la surveillance post-commercialisation, les procédures de signalement d'incidents graves, et la gestion de la communication avec les autorités compétentes. Pour les organisations disposant déjà d'une certification **ISO 27001** (sécurité de l'information) ou **ISO 42001** (système de management de l'IA), l'intégration des exigences de l'AI Act dans le SMQ existant peut se faire de manière incrémentale. Les normes harmonisées européennes, en cours d'élaboration par le CEN et le CENELEC, fourniront à terme des référentiels techniques détaillés permettant de présumer la conformité aux exigences de l'AI Act.

Surveillance Post-Commercialisation (Article 72)

La conformité ne s'arrête pas à la mise sur le marché : l'article 72 impose une **surveillance post-commercialisation** proportionnée à la nature du système IA et aux risques identifiés. Le fournisseur doit établir et documenter un système de surveillance post-commercialisation intégré dans son SMQ. Ce système doit permettre de collecter, documenter et analyser activement les données pertinentes fournies par les déployeurs ou collectées via d'autres sources, afin d'évaluer en continu la conformité du système aux exigences du règlement tout au long de sa durée de vie. Pour les LLM, cette surveillance inclut le monitoring des performances du modèle en production (détection de la **dérive du modèle** ou model drift), l'analyse des retours utilisateurs et des plaintes, la détection des comportements inattendus ou dangereux, et la veille sur les nouvelles vulnérabilités et techniques d'attaque adversariale. Les incidents graves — définis comme tout incident entraînant directement ou indirectement un décès, un dommage grave à la santé, une violation grave des droits fondamentaux ou un dommage grave aux biens, à l'environnement ou aux infrastructures critiques — doivent être signalés aux autorités de surveillance dans un délai de **15 jours** suivant leur identification. Pour approfondir, consultez Mixture of Experts (MoE) : Architecture, Sécurité et.



Obligations Haut Risque Conformité Pratique Roadmap Conformité



7 Roadmap de Mise en Conformité AI Act

La mise en conformité avec l'AI Act est un **programme pluriannuel** qui doit être structuré en phases cohérentes, alignées sur le calendrier progressif d'entrée en application du règlement. En février 2026, les organisations se trouvent dans une fenêtre stratégique critique : les interdictions de l'article 5 sont déjà en vigueur, les obligations GPAI entreront en application dans six mois, et les exigences relatives aux systèmes à haut risque suivront un an plus tard. La roadmap suivante propose une approche structurée en quatre phases pour les organisations déployant des systèmes LLM.

Phase 1 : Inventaire et Cartographie (T1-T2 2026)

La première phase, à engager immédiatement, consiste à réaliser un **inventaire exhaustif de tous les systèmes IA** déployés ou en cours de développement au sein de l'organisation. Cet inventaire doit couvrir les systèmes développés en interne, les solutions SaaS intégrant des composants IA, les modèles GPAI utilisés via API (OpenAI, Anthropic, Google, Mistral), et les usages informels de l'IA par les collaborateurs (shadow AI). Pour chaque système

identifié, documentez la finalité prévue, les données traitées, les personnes affectées, le fournisseur du modèle sous-jacent, et le niveau de risque préliminaire selon la pyramide de l'AI Act. Constituez un **registre centralisé des systèmes IA** qui servira de base à toute la démarche de conformité. Identifiez le responsable métier et le responsable technique de chaque système. Cette phase doit également inclure une analyse des pratiques interdites de l'article 5 pour vérifier qu'aucun système existant ne tombe dans cette catégorie, car ces interdictions sont déjà applicables et les sanctions immédiates.

Phase 2 : Classification et Analyse d'Écarts (T2-T3 2026)

La deuxième phase consiste à **classifier formellement chaque système** selon les quatre niveaux de risque de l'AI Act et à réaliser une analyse d'écarts (gap analysis) entre la situation actuelle et les exigences réglementaires applicables. Pour chaque système classé à haut risque, évaluez le niveau de maturité actuel par rapport aux neuf obligations principales : système de gestion des risques (art. 9), gouvernance des données (art. 10), documentation technique (art. 11), journalisation (art. 12), transparence et information (art. 13), surveillance humaine (art. 14), exactitude, robustesse et cybersécurité (art. 15), système de management de la qualité (art. 17), et évaluation de conformité (art. 43). Quantifiez les écarts identifiés en termes d'**effort de mise en conformité** (en jours-homme), de coûts estimés et de délais prévisionnels. Priorisez les actions en fonction des dates d'entrée en application et de la criticité des systèmes concernés. Cette phase est l'occasion de déterminer si certains systèmes doivent être redessinés, abandonnés ou remplacés par des alternatives moins risquées.

Phase 3 : Mise en Conformité Opérationnelle (T3 2026 - T2 2027)

La troisième phase est la plus intensive : elle consiste à **implémenter concrètement les mesures de conformité** identifiées lors de l'analyse d'écarts. Pour les systèmes à risque limité (chatbots, générateurs de contenu), les actions prioritaires sont la mise en œuvre de notifications de transparence conformes à l'article 50, l'implémentation du marquage des contenus synthétiques, et la documentation des mécanismes de détection d'IA. Pour les systèmes à haut risque, le programme de mise en conformité est beaucoup plus structurant. Il inclut le déploiement d'une infrastructure de **logging et de monitoring** conforme aux exigences de l'article 12, la mise en œuvre de pipelines de test automatisés pour l'évaluation continue des biais, de la robustesse et de la cybersécurité, la rédaction de la documentation technique complète conformément à l'Annexe IV, la formation des opérateurs humains chargés de la surveillance du système, la mise en œuvre du système de gestion des risques itératif, et la préparation de l'évaluation de conformité (auto-évaluation ou organisme notifié selon le cas). Parallèlement, intégrez les exigences GPAI dans vos contrats avec les fournisseurs de modèles en demandant la documentation technique, le résumé des données d'entraînement et la politique de droit d'auteur.

Phase 4 : Gouvernance Continue et Amélioration (T2 2027+)

La quatrième phase installe un régime de **gouvernance continue** de la conformité IA, conçu pour durer au-delà de la mise en conformité initiale. Mettez en place un **comité de gouvernance IA** regroupant les fonctions juridique, technique, métier, conformité, DPO et

cybersécurité, chargé de superviser l'ensemble du portefeuille de systèmes IA de l'organisation. Définissez un processus de revue périodique (au minimum semestrielle) de chaque système à haut risque, intégrant l'analyse des données de monitoring, les retours utilisateurs, les incidents signalés et l'évolution du contexte réglementaire. Implémentez un processus de **gestion du changement** pour tout nouveau déploiement ou toute modification significative d'un système IA existant, incluant une évaluation de classification et une analyse d'impact préalables. Préparez-vous aux **audits** des autorités de surveillance nationales (en France, la CNIL en coordination avec l'autorité de surveillance IA à désigner) en maintenant à jour l'ensemble de la documentation technique et du registre des systèmes IA. Enfin, participez activement aux travaux de normalisation (CEN/CENELEC) et aux consultations de l'AI Office pour anticiper l'évolution des exigences techniques et des bonnes pratiques reconnues par les autorités européennes.

Points clés de la roadmap :

Pour approfondir ce sujet, consultez notre outil open-source ai-threat-detection qui facilite la détection de menaces basée sur l'IA.

- ▶**Février 2025 :** Interdictions art. 5 en vigueur — vérifiez immédiatement vos systèmes existants
- ▶**Août 2025 :** Obligations GPAI — exigez la documentation de vos fournisseurs de modèles
- ▶**Août 2026 :** Systèmes haut risque Annexe III — finalisez votre évaluation de conformité
- ▶**Août 2027 :** Systèmes haut risque Annexe I — conformité produits réglementés
- ▶**2028+ :** Gouvernance continue, audits, amélioration permanente du SMQ IA

Ressources open source associées

HF Space ai-act-risk-classifier (démo) HF Dataset ai-act-fr

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
- EUR-Lex — AI Act — Règlement européen sur l'intelligence artificielle

Peut-on contester la classification d'un systeme IA par l'AI Act ?

Oui, les fournisseurs de systemes IA peuvent contester leur classification aupres des autorites nationales competentes. Le reglement prevoit des mecanismes de recours et de revision. Il est recommande de documenter rigoureusement l'analyse de risque pour justifier le niveau de classification retenu.

Combien coute la mise en conformite AI Act pour une PME ?

Le cout de mise en conformite AI Act varie significativement selon le type de systeme IA et son niveau de risque. Pour une PME avec un systeme a risque limite, le budget se situe entre 15 000 et 50 000 euros, incluant l'audit, la documentation technique et la formation des equipes.

Quels sont les prérequis techniques pour déployer AI Act et LLM : Classifier vos Systèmes IA ?

Il faut un environnement Python 3.10+, des GPU compatibles CUDA si vous traitez de gros volumes, et un accès aux API des modèles utilisés. Prévoyez aussi un pipeline de données propre et documenté.

Sources et références : ArXiv IA · Hugging Face Papers

Conclusion

Cet article a couvert les aspects essentiels de Table des Matières, 1 L'AI Act : Le Règlement Européen sur l'Intelligence Artificielle, 2 La Pyramide des Risques : 4 Niveaux de Classification. 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