Opérationnaliser l'AI Act : Gouvernance, XAI et ingénierie de conformité des systèmes d'IA

Septembre 2026

Abstract

Cet article propose une analyse approfondie du règlement européen sur l'intelligence artificielle (AI Act), en postulant que ce nouveau cadre légal, parfois présenté comme un carcan réglementaire, constitue en réalité un recueil de bonnes pratiques essentielles pour le développement de systèmes robustes et performants. L'étude débute par une mise en contexte technologique des modèles d'IA, tout en soulignant les risques inhérents à leur autonomie croissante, tels que l'opacité, les cyberattaques et les biais latents. Le document présente ensuite la taxinomie des risques retenue par le régulateur, puis cartographie la gouvernance impliquant instances européennes et nationales (comme la CNIL ou l'ARCOM en France). L'opérationnalisation des exigences de l'AI Act est discutée en s'appuyant sur le triptyque FAT (*Fairness, Accountability, Transparency*). L'auteur détaille comment le principe d'équité impose une rigueur sur la qualité et la représentativité des données d'entraînement, incluant une dérogation pour le traitement de données sensibles afin de corriger les biais éventuels. Le principe de responsabilité exige quant à lui une gestion stricte des risques, une supervision humaine continue et une traçabilité par journalisation des événements. En matière de transparence, la réglementation requiert des notices d'utilisation claires et le marquage univoque des contenus générés. Une attention particulière est accordée à l'intelligence artificielle explicable (XAI), vitale pour maîtriser les risques opérationnels. L'auteur présente des méthodes d'explicabilité locale (SHAP, LIME, gradients locaux) et globale (modèles de substitution, Concept Activation Vectors), tout en insistant sur la nécessité d'évaluer la fidélité, la stabilité et la compréhensibilité de ces explications. L'article explore également l'ingénierie de la conformité, soulignant l'importance d'outils tels que MLflow pour l'auditabilité et Fairlearn pour la mesure de l'équité. Enfin, l'étude démontre que l'AI Act s'intègre harmonieusement dans les cadres réglementaires sectoriels préexistants, tels que Bâle III pour la finance ou le RDM pour la santé, en unifiant gouvernance des données, supervision des modèles et piste d'audit. En définitive, ce document offre un décryptage technique et juridique montrant que l'anticipation de ces normes est un vecteur de création de valeur à long terme pour les entreprises. Note : Cet article a été rédigé avec l'assistance de Gemini.

Mots clefs : EU AI Act, Risk-based approach, Framework FAT, XAI, GPAIM, MLOps, Auditability, Algorithmic Bias, Data Quality

DATA & AI Empowerment

Atheïa - Sas au capital de 50 000 € - 881 048 938 RCS Paris
17, rue de Choiseul – 75 002
Paris
881 048 938 R.C.S. Paris

www.atheia.fr

1 Introduction

L'intégration massive des technologies d'intelligence artificielle (IA) au sein de processus industriels critiques exige, pour nombre d'entreprises, de renforcer leur cadre de gouvernance et de gestion des risques. L'entrée en vigueur de l'IA Act européen (Règlement UE 2024/1689) [1] et les ajustements apportés à l'été 2026 par le régulateur (« Digital Omnibus ») [2] redessinent en profondeur les exigences applicables aux entreprises européennes fournissant, déployant ou utilisant des systèmes d'IA.

Pour l'auteur de cet article, ce nouveau cadre réglementaire, loin de constituer une surcouche de contraintes bureaucratiques, recouvre un ensemble de bonnes pratiques visant la mise en place d'architectures d'IA plus robustes, plus performantes et créatrices de valeur à long terme. En raison de la récence du sujet, la littérature de vulgarisation reste confidentielle. Il nous a donc semblé pertinent de proposer un décryptage du nouveau paradigme réglementaire et technique de l'IA européenne à destination des professionnels susceptibles de mettre en œuvre des projets d'IA :

- Objectif 1. Présenter les bonnes pratiques industrielles en matière de développement et gouvernance de projet analytique et informatique, tout au long du cycle de vie du produit.
- Objectif 2. Démystifier l'AI Act européen en le rattachant à des considérations concrètes et des logiques industrielles déjà appropriées par les professionnels du secteur.
- Objectif 3. Regrouper des ressources techniques et méthodologiques utiles aux professionnels impliqués dans l'opérationnalisation de l'IA.

Le reste de cet article est organisé comme suit. Nous proposons en section 2 un bref rappel des différents modèles d'IA, des risques liés à leur utilisation, ainsi qu'une présentation des différents acteurs du nouveau cadre réglementaire. Puis nous discutons des principes fondateurs d'une IA saine et robuste, en analysant les exigences de l'AI Act au prisme du triptyque FAT (*Fairness, Accountability, Transparency*) dans la section 3. Nous poursuivons l'analyse en section 4 en proposant un focus sur les méthodes d'IA explicable (XAI). Nous livrons dans la section 5 une boite à outils regroupant un ensemble de ressources réglementaires et techniques pertinentes pour une entreprise désireuse de développer ou déployer un système d'IA. Enfin, l'adhérence entre l'AI Act et les réglementations sectorielles existantes est discutée en section 6.

2 L'écosystème de l'IA européenne

2.1 L'IA, de quoi parle-t-on ?

Pour appréhender la portée de la régulation, il convient d'abord d'en définir l'objet technique. L'AI Act européen, complété par les orientations de la Commission [3], propose une définition de l'intelligence artificielle technologiquement neutre, c'est-à-dire davantage liée à son fonctionnement qu'à sa conception. Elle est aussi plastique, s'appuyant principalement sur la notion d'inférence, anticipant ses potentielles évolutions futures. Selon l'Article 3, un système d'IA (SIA) est défini comme un système automatisé conçu pour fonctionner à différents niveaux d'autonomie, capable de déduire (inférer) des modèles à partir de données d'entrée pour générer des sorties (ex : prédictions [SIC],

recommandations, contenus, décisions) susceptibles d'influencer des environnements physiques ou virtuels.

D'un point de vue technologique, nous pouvons lister une large gamme de modèles constitutifs d'un écosystème de l'IA (ex : *computer vision*, *speech recognition*, *robotics*, *NLP*). Néanmoins, nous avons choisi pour cette étude de mettre en lumière deux grandes familles de modèles :

- **L'IA prédictive et discriminative traditionnelle** : Fondée sur des techniques éprouvées de *machine learning* et de *deep learning*, elle vise à classer, scorer ou prévoir à partir de données historiques hautement structurées. Ces modèles (ex : régressions, modèles d'arbres, réseaux de neurones convolutionnels) s'appuient principalement sur le paradigme du *fine-tuning* traditionnel, où un modèle pré-entraîné ou initialisé aléatoirement est ajusté (paramètres et hyperparamètres) sur un jeu de données étiqueté spécifique à une tâche en aval.
- **L'IA générative et les modèles d'IA à usage général ou *General Purpose AI Models* (GPAIM) [4]** : Portée par la rupture technologique des architectures de type *Transformer*, cette famille englobe les grands modèles de langage (LLM). Ces modèles se caractérisent par une généralité significative et une capacité à exécuter un large spectre de tâches distinctes. Leur fonctionnement repose le plus souvent sur un paradigme de requêtage (*prompting*) : ils subissent d'abord un pré-entraînement auto-supervisé sur un large corpus textuel pour générer des modèles de base (*base models*), puis sont spécialisés sur une tâche ou un domaine via un sur-apprentissage (ou affinage) spécifique supervisé (SFT). Ils sont à ce stade capables d'interagir en langage naturel. Enfin, leurs capacités de raisonnement sont augmentées et leur comportement aligné avec les préférences humaines via des techniques d'apprentissage par renforcement (ex : RLHF) ou via des approches d'optimisation par récompense (ex : DPO, RGPO, RLVR).

2.2 Pourquoi réguler ?

L'autonomie croissante des systèmes d'IA engendre d'importants risques opérationnels lorsqu'ils sont déployés dans des environnements critiques. Les risques les plus saillants incluent l'opacité des modèles complexes de type « boîte noire », l'exfiltration involontaire de données confidentielles, les cyberattaques par empoisonnement de données (*data poisoning*) ou par exemples antagonistes (*adversarial attacks*), ainsi que la génération d'hallucinations ou de biais discriminatoires encodés de façon latente dans les jeux de données d'entraînement.

Pour prévenir et atténuer ces risques, la Commission a structuré la régulation de l'AI Act européen autour d'une approche proportionnée fondée sur une taxinomie de ces risques, applicable à toute entreprise ou agence :

- **Risque inacceptable (Pratiques interdites) [5]** : Sont formellement interdits depuis février 2025 les systèmes d'IA déployant des techniques de manipulation subliminale causant un préjudice aux utilisateurs, notamment la notation sociale (*social scoring*) par des acteurs publics ou privés, la catégorisation biométrique basée sur des attributs sensibles, ou encore l'analyse des émotions sur le lieu de travail par exemple.
- **Haut risque (*High-Risk AI Systems*)** : Soumis à un cadre de conformité rigoureux, ce groupe comprend les systèmes listés à l'Annexe III (ex : tri automatisé de CV, l'évaluation de la

solvabilité des personnes physiques, la gestion et l'exploitation des infrastructures critiques d'approvisionnement en eau, électricité, gaz) ainsi que les systèmes d'IA intégrés comme composants de sécurité dans des produits réglementés par l'Annexe I (ex : dispositifs médicaux, machines industrielles).

- **Risque limité (Obligations de transparence)** : Concerne les systèmes interagissant avec l'humain (*chatbots*), la catégorisation biométrique ou la génération de contenus synthétiques (*deepfakes*). En dépit d'obligations allégées, l'Article 50 impose l'intégration de marquages lisibles par machine (filigranes ou *watermarking*) et une divulgation claire d'une interaction avec un système ou un modèle d'IA.
- **Risque minimal** : Applications standards (ex : *spam filters*, outils marketing) exemptées d'obligations, mais encouragées à adopter des chartes éthiques volontaires.

Précisons que ces deux dernières catégories sont utilisées par la Commission européenne dans ses communications, mais ne sont pas reprises dans le règlement en tant que tel. Une liste d'exclusion est introduite à l'Article 2.

2.3 Gouvernance européenne et nationale, qui sont les acteurs ?

L'AI Act introduit une architecture de gouvernance multi-niveaux pour superviser le marché unique européen. Au niveau de l'Union européenne, son application est coordonnée par le **Bureau européen de l'intelligence artificielle,** rattaché à la Commission européenne. Il est appuyé par le **Comité européen de l'IA**, réunissant des représentants des États membres pour favoriser l'harmoniser des pratiques, et par un **panel scientifique d'experts indépendants** chargé d'alerter sur les risques systémiques des modèles d'IA à usage général.

Au niveau national, les États membres désignent des autorités de surveillance du marché. En France, une vingtaine d'agences nationales se partagent la supervision par secteur d'activité, parmi lesquelles la CNIL[1] et l'ARCOM[2] couvrent les plus nombreux cas d'usage. La DGCCRF[3] assure la coordination opérationnelle de ces autorités, tandis que la DGE[4] assure la coordination stratégique et représente la France au Comité de l'IA. Enfin, l'ANSSI[5] et le PEReN[6] appuient les autorités nationales dans leurs missions de contrôle de la conformité en offrant un socle de compétences techniques mutualisé.

Par ailleurs, le règlement structure les responsabilités en distinguant les rôles au long de la chaîne de valeur :

- Le **Fournisseur** : Personne physique ou morale qui développe un système d'IA ou le fait développer, en vue de le mettre sur le marché ou de le mettre en service sous son propre nom ou sa marque. Il porte la responsabilité majeure de la conformité technique.

[1] Commission nationale de l'informatique et des libertés.
[2] Autorité de régulation de la communication audiovisuelle et numérique.
[3] Direction générale de la concurrence, de la consommation et de la répression des fraudes.
[4] Direction générale des Entreprises.
[5] Agence nationale de la sécurité des systèmes d'information.
[6] Pôle d'expertise de la régulation numérique.

- Le **Déployeur** : Toute entité professionnelle utilisant un système d'IA sous sa propre autorité. Ses obligations portent sur le respect des notices d'utilisation, la supervision humaine locale, la journalisation et la réalisation d'analyses d'impact sur les droits fondamentaux.

La distinction entre fournisseur et déployeur n'est pas toujours aisée. Par exemple, un déployeur apportant une « modification substantielle » à un système d'IA à haut risque, ou modifiant la destination d'un système d'IA de telle sorte qu'il devient à haut risque, en devient le **fournisseur,** endossant l'intégralité des obligations de conformité. Nous voyons ici une zone grise dans l'application des définitions de « modification substantielle » et « destination » qui appelleront sûrement une clarification du régulateur dans les mois qui viennent. Le contexte étant posé, nous abordons le premier objectif de notre étude en discutant des bonnes pratiques industrielles pour la construction d'un système d'IA sain.

3 Principes fondateurs d'une IA saine

Bâtir un système d'IA sain exige de définir un cadre d'éthique et de transparence. Le triptyque de l'approche **FAT (*Fairness, Accountability, Transparency*)** peut justement être mobilisé pour opérationnaliser les sept principes clés édictés par le Groupe d'experts de haut niveau [6] et repris par la Commission.

Table 1. Principes éthiques et framework FAT.

Thème	Principe
Équité	▪ Diversité, non-discrimination et équité ▪ Bien-être sociétal et environnemental
Responsabilité	▪ Robustesse et sécurité ▪ Facteur humain et contrôle humain ▪ Responsabilisation
Transparence	▪ Respect de la vie privée et gouvernance des données ▪ Transparence

Dans cette section, nous verrons comment l'AI Act européen reprend un certain nombre de ces dispositions, au titre des obligations incombant aux fournisseurs et déployeurs de SIA à haut risque ou risque limité. La section 4, quant à elle, vise à expliciter ces dispositions en s'appuyant sur les méthodes et outils d'une IA explicable.

3.1 Principe d'équité

Le principe d'équité vise à prévenir la création ou la perpétuation de biais pouvant porter préjudice à une population. Une attention particulière doit être portée à la collecte des données avant et après la mise en service du SIA :

- **Qualité des données** : Les jeux de données d'entraînement, de validation et de test doivent être « pertinents, représentatifs et, dans la mesure du possible, exempts d'erreurs et complets ». Ils doivent posséder des propriétés statistiques appropriées en rapport avec la destination du modèle, telles que la représentativité correcte des sous-groupes démographiques afin d'éviter les biais de sous-représentation.
- **Dérogation concernant les données sensibles** : De manière exceptionnelle, et sous réserve de garanties de pseudonymisation et de sécurité strictes, l'Article 10 autorise le traitement de données sensibles (ex : catégories particulières de données au sens de l'Article 9 du RGPD, telles que l'ethnie, la religion ou l'orientation sexuelle) pour l'unique objectif de détecter et corriger des biais discriminatoires reproduits par le modèle. Ces données doivent être détruites dès la correction du biais.
- **Analyse d'Impact sur les Droits Fondamentaux** : La cartographie des conséquences d'un déploiement d'un SIA sur les populations vulnérables et sa mise à jour après implémentation comptent au nombre des bonnes pratiques de gestion des SIA. De cette cartographie sont élaborés les plans de remédiation visant à atténuer, voire éviter les impacts négatifs sur ces populations vulnérables, dont les résultats sont notifiés à l'autorité nationale de surveillance.
- **Frugalité énergétique** : Si une conception économe en ressources et une planification des cycles de réentraînement à une fréquence raisonnable sont encouragées pour limiter l'empreinte carbone du traitement informatique, la question de la transition écologique est évacuée de la réglementation. Le régulateur reste fidèle à la doctrine développée jusqu'ici, consistant à prêter attention aux risques que le climat fait porter à la sphère économique, et non l'inverse (une vingtaine d'occurrence du mot « économique », une seule pour « écologique »).

3.2 Principe de responsabilité

Le principe de responsabilité repose sur une traçabilité des opérations et des décisions. Une gouvernance performante implique la mise en place d'un processus itératif et documenté sur l'intégralité du cycle de vie du modèle :

- **Système de gestion des Risques** : Un SIA sain doit établir un système robuste de gestion des risques. Ce système doit permettre l'identification des risques connus et raisonnablement prévisibles pour la santé, la sécurité ou les droits fondamentaux, l'estimation des risques liés à une « mauvaise utilisation raisonnablement prévisible » et la planification des mesures d'atténuation fondées sur l'état de l'art technique.
- **Gestion des incidents** : Tout dysfonctionnement ou incident grave, comme ceux portant atteinte aux droits fondamentaux ou à l'intégrité physique, doit être documenté et notifié dans les plus brefs délais aux autorités nationales de surveillance de marché.
- **Supervision Humaine** : Les systèmes doivent être conçus pour permettre à des personnes physiques de surveiller leur fonctionnement, de détecter les anomalies et d'interrompre ou d'invalider (*override*) une recommandation algorithmique erronée. Les équipes en charge de cette supervision doivent être formées et disposer de l'autorité hiérarchique nécessaire.
- **Traçabilité par journalisation** : À des fins d'audit et d'investigation, la conception des SIA doit intégrer l'enregistrement automatique d'événements (*logs*) ainsi que leur archivage.

Ceci afin d'identifier les situations présentant un risque, retracer les prévisions (ex : bases de données interrogées, filtres appliqués), ou établir l'identité de l'opérateur humain par exemple.

- **Audit et contrôle** : La gouvernance d'un SIA sain doit permettre de définir des responsabilités claires en première, deuxième et troisième lignes de défense, et d'offrir un cadre auditable lors d'inspections des autorités de supervision.

3.3 Principe de transparence

Enfin, le principe de transparence vise à restituer à l'humain la compréhension et le contrôle des décisions du système et de la provenance des données :

- **Notice d'utilisation** : Les fournisseurs de SIA doivent mettre à disposition une notice précisant la destination du système, ses capacités, ses limites, les spécifications liées aux données d'entrée attendues et les mesures participant au contrôle humain.
- **Marquage de l'IA Générative** : Par souci de transparence, les utilisateurs doivent être informés de façon univoque lorsqu'ils interagissent directement avec un SIA. De la même façon, les contenus résultant d'un SIA (ex : texte, image, audio) doivent être marqués d'un étiquetage détectable par machine, certifiant leur origine artificielle.
- **Conformité RGPD et lignes directrices de la CNIL** : Le développement de modèles d'IA sur données personnelles doit impérativement définir une finalité précise et légitime. L'exercice des droits individuels (ex : accès, rectification, effacement) doit être nativement intégré dans l'architecture de données, tout comme les pratiques d'anonymisation ou de pseudonymisation des données au sein du jeu historique d'apprentissage. Ceci afin de prévenir le coût disproportionné du réentraînement complet d'un modèle d'IA pour répondre à une demande d'effacement ou d'opposition par exemple.

4 Explainable Artificial Intelligence (XAI)

L'opacité inhérente aux modèles d'IA, souvent qualifiés de « boîtes noires », engendre des risques opérationnels majeurs, tels que l'amplification des biais algorithmiques, la dilution de la responsabilité juridique, ou l'érosion de la confiance des parties prenantes. Le champ de l'intelligence artificielle explicable (XAI - *Explainable Artificial Intelligence*) vise justement à atténuer ces risques. Celui-ci repose sur le triptyque FAT évoqué précédemment comprenant **transparence** (soit l'intelligibilité des sorties pour un être humain), **responsabilité** (la capacité à reconstituer la chaîne causale des données) et **équité** (la légitimité éthique et technique de la décision).

Loin de représenter une simple exigence déclarative de conformité, elle constitue un levier de performance basé sur un ensemble d'outils techniques indispensables à la maîtrise des risques opérationnels. Concrètement, il s'agit de structurer l'analyse du fonctionnement du modèle en intégrant des indicateurs locaux et globaux, puis de soumettre les explications générées à une évaluation rigoureuse de leur qualité intrinsèque.

4.1 Méthodes d'explicabilité locale

L'explicabilité locale a pour objectif d'éclairer l'analyste sur le comportement d'un modèle d'IA au niveau d'une instance ou d'une prévision individuelle. Elle participe des principes de responsabilité (traçabilité des opérations et des décisions) et de transparence (compréhension et contrôle des décisions du système) évoqués précédemment. Concrètement, elle cherche à répondre à la question suivante : quelles caractéristiques d'entrée spécifiques ont déterminé cette sortie particulière ?

Modèles prédictifs traditionnels

Dans le cadre de l'apprentissage supervisé traditionnel, l'explicabilité locale repose principalement sur des méthodes d'attribution de caractéristiques (*feature attribution*) *post hoc* et agnostiques au modèle. Les méthodes SHAP [7] et LIME [8] sont largement utilisées bien qu'elles peinent à satisfaire aux critères de stabilité et fidélité évoqués dans la section 4.3.

L'approche **SHAP** (*Shapley Additive exPlanations*), qui repose sur les valeurs de Shapley, évalue la contribution marginale moyenne de chaque variable d'entrée à la décision finale, à travers l'ensemble des combinaisons possibles de caractéristiques.

Exemple d'application : Dans un système d'octroi de crédit bancaire, SHAP permet de justifier le rejet d'une demande individuelle en estimant que le ratio d'endettement a contribué positivement au score de risque à hauteur de 2,5 tandis que l'ancienneté professionnelle n'a réduit ce risque que de 0,05.

Pour expliquer une décision individuelle selon l'approche **LIME** (*Local Interpretable Model-Agnostic Explanations*), l'instance d'entrée est perturbée dans son voisinage immédiat, puis un modèle de substitution local nativement interprétable (ex : régression linéaire, arbre de décision) est ajusté sur les variations de sorties du modèle principal.

Exemple d'application : Pour un classifieur de diagnostic d'imagerie médicale, LIME génère des perturbations en masquant des segments de l'image (super-pixels) pour identifier la région anatomique qui a guidé la classification.

Modèles génératifs et grands modèles de langage

Le nombre de paramètres des modèles à usage général induit une complexité calculatoire qui limite l'usage direct des méthodes de perturbation classiques en temps réel. Plusieurs approches ont été développées pour contourner cette limitation.

L'attribution par gradients locaux évalue la sensibilité locale en calculant les dérivées partielles de la fonction de prévision (ex : probabilité du token suivant) par rapport aux dimensions de l'espace des représentations vectorielles d'entrée.

Exemple d'application : Dans une tâche de diagnostic (ex : étude de conformité), la méthode du gradient permet de mesurer et cartographier la contribution de chaque mot du texte analysé (ex : clause contractuelle spécifique) dans la réponse générée par le LLM assistant.

L'utilisation de **mécanismes d'attention hybrides** (*Function-Based attention*) combine la sensibilité des sorties avec les poids d'attention pour localiser précisément les interactions d'informations dans les couches du Transformer.

Exemple d'application : Dans une tâche d'analyse de sentiment de commentaires clients, un indicateur hybride attention-gradient met en évidence que l'association sémantique entre le sujet « service » et l'adjectif « déplorable » a dicté la classification négative.

4.2 Méthodes d'explicabilité globale

L'explicabilité globale vise à offrir une compréhension macroscopique du fonctionnement du modèle sur l'ensemble de son espace d'apprentissage. Elle permet d'auditer la cohérence des représentations apprises et de valider l'absence de dérives systémiques. En cela, elle s'inscrit principalement dans le principe de responsabilité, mais aussi dans ceux de transparence (identification des capacités et des limites) et d'équité en fournissant les outils nécessaires à une étude d'impact sur les droits fondamentaux.

Modèles prédictifs traditionnels

Ces modèles constituent la référence d'interprétabilité globale directe, car leur structure interne reflète une logique intelligible, sous réserve d'un nombre de paramètres humainement appréhendable. Les modèles de régression offrent une forme explicite incluant le poids de chaque paramètre, quand le chemin de prévision (*prediction path*) est lisible nativement pour les modèles d'arbre.

Lorsque le nombre de paramètres augmente au point d'empêcher toute lecture directe (ex : modèles hautement non linéaires), une **distillation de modèle** peut être envisagée afin de réduire le nombre de paramètres tout en conservant le comportement du modèle d'origine. Ce modèle plus simple, dit de **substitution** (*Global Surrogates*), peut être entraîné pour imiter au mieux les frontières de décision du modèle complexe et faciliter son interprétation.

Exemple d'application : En tarification d'assurance, un modèle de forêt aléatoire (*random forest*) complexe peut être approximé globalement par un modèle linéaire de substitution afin de fournir aux régulateurs un tableau de coefficients globaux stables et auditables.

Modèles génératifs et grands modèles de langage

Interpréter globalement un modèle comptant plusieurs dizaines de milliards de paramètres nécessite de développer des approches alternatives propres à cartographier des concepts encodés dans les composants individuels (ex : neurones, couches).

Les méthodes ***Classifier-Based Probing*** visent à mesurer la qualité de l'encodage des propriétés linguistiques complexes, telles que la syntaxe ou la sémantique, dans les couches d'activation du modèle. Ces méthodes s'appuient sur un modèle simple (*Probe Classifier*), entraîné en surcouche du modèle de base afin de détecter les propriétés linguistiques d'intérêt. La qualité de cet apprentissage complémentaire dépend directement des propriétés linguistiques encodées dans le modèle de base. Empiriquement, les propriétés syntaxiques sont davantage encodées dans les premières couches alors que les propriétés sémantiques le sont dans les dernières.

Plutôt que d'analyser des propriétés sémantiques, la méthode **CAV** (*Concept Activation Vectors*) [9] permet d'identifier des concepts de haut niveau définis par le testeur. Un classifier linéaire est entraîné

à identifier les activations neuronales d'un concept au travers un jeu de données d'illustration (ex : le concept de « neutralité » ou de « sécurité ») par rapport à un ensemble aléatoire. L'analyse par le classifier des concepts encodés les plus utilisés pour générer la prévision permet de quantifier la sensibilité globale du modèle à ce concept spécifique.

Exemple d'application : Dans l'analyse globale d'un LLM appliqué à la santé, CAV permet de vérifier dans quelle mesure le concept « symptômes cardiovasculaires » influence positivement le diagnostic global du modèle par rapport à d'autres concepts-bruits.

4.3 Évaluer la qualité des explications

Disposer de méthodes d'explicabilité est une condition nécessaire mais insuffisante. Il convient d'évaluer si les explications fournies sont **fidèles** au fonctionnement du modèle, **stables**, et réellement **compréhensibles** pour un opérateur humain, en vertu du principe de transparence.

Le principe de **fidélité** (*Fidelity*) mesure l'exactitude avec laquelle l'explication reflète la véritable logique décisionnelle du modèle étudié (ex : modèle « boîte noire »). Deux indicateurs complémentaires [10] peuvent être utilisés pour détecter et éliminer les rationalisations *post hoc* trompeuses ou indépendantes de la causalité réelle :

- La **compréhensivité** (*Comprehensiveness*) mesure la baisse de probabilité de la classe prédite (ou du token) lorsque l'on retire de l'entrée les caractéristiques identifiées par l'explication comme les plus importantes.
- La **suffisance** (*Sufficiency*) évalue si les caractéristiques identifiées par l'explication comme les plus importantes se suffisent à elles-mêmes pour que le modèle conserve une probabilité de prédiction similaire à celle obtenue avec l'entrée complète.

Une compréhensivité élevée et une suffisance faible signalent une explication fidèle.

La **stabilité** (*Stability*) garantit que des instances d'entrée très proches reçoivent des explications similaires, éliminant ainsi les explications erratiques générées par du bruit ou évitant la situation où deux cas d'usage identiques reçoivent des justifications radicalement différentes.

Enfin, le principe de **compréhensibilité** (*Comprehensibility*) propose d'évaluer la charge cognitive requise par un humain pour analyser l'explication fournie, en adaptant le format au destinataire (ex : explications visuelles pour les développeurs, explications textuelles simplifiées pour les régulateurs ou usagers), ou le nombre de caractéristiques retenues pour l'explication.

Cet article s'adressant principalement aux praticiens qui ne peuvent pas toujours s'offrir le temps d'une revue bibliographique, nous avons sélectionné deux articles scientifiques particulièrement éclairants pour notre propos :

- Bamigbade et al. (2024) [11] examinent la convergence des principes d'éthique et d'explicabilité dans le développement de l'IA, en proposant un cadre d'analyse pour leur intégration dans le cycle de vie d'un modèle.
- Zhao et al. (2024) [12] proposent une étude structurée des méthodes d'explicabilité des larges modèles de langage fondés sur les *Transformers*.

5 Comment construire son système d'IA ?

Les principes introduits aux section précédentes ne sont pas que pure théorie des systèmes d'information. Un responsable R&D et un data scientist devront en effet savoir les traduire en principes de gouvernance, processus métier ou lignes de code. Cette opération exige de se poser les bonnes questions et d'adopter les bons outils. Nous proposons dans cette section un ensemble de ressources pertinentes pour le praticien.

5.1 Les bonnes questions à se poser

Avant toute mise en production, voire pendant la phase de cadrage, chaque acteur de la chaîne de valeur doit confronter son projet à une série de questions précises. Ci-dessous une liste non-exhaustive, alignées sur le cadre d'analyse FAT :

- **Le Fournisseur (R&D / Dev)** :
  - *Mon cas d'usage relève-t-il du haut risque au sens de l'Annexe III (ex : recrutement, scoring de crédit) ?*
  - *Mes jeux de données de test sont-ils statistiquement représentatifs de ma population cible (équité) ?*
  - *Puis-je prouver formellement le lignage de chaque donnée utilisée pour mon apprentissage (responsabilité) ?*
  - *Dois-je intégrer des facteurs d'explicabilité spécifiques ou agnostiques au modèle implémenté (responsabilité) ?*
- **Le Déployeur (Directeur Métier / Intégrateur)** :
  - *Ai-je configuré les pipelines d'archivage des logs pour qu'ils conservent les traces de décision pendant au moins six mois de façon immuable (responsabilité) ?*
  - *Le personnel désigné pour superviser l'outil a-t-il reçu une formation adéquate à la littératie de l'IA pour déceler les biais d'automatisation (transparence) ?*
  - *L'outil permet-il à un opérateur d'annuler ou de modifier manuellement la recommandation algorithmique (responsabilité) ?*

Enfin, même si l'utilisateur final ne participe pas au développement ni au déploiement du SIA, il reste un acteur essentiel, dont l'intérêt mérite d'être considéré par les autres parties prenantes :

- **L'Utilisateur final** :
  - *Le système affiche-t-il une information claire lui indiquant qu'il interagit avec une machine (IHM de* chatbot*) (transparence) ?*
  - *Peut-il formuler un recours ou obtenir une explication intelligible sur le critère précis qui a motivé son score (transparence) ?*

5.2 Notre boite à outils

Les questions évoquées plus haut invitent au diagnostic technique visant à choisir le meilleur outil pour une tâche donnée. L'ingénierie de la conformité propose de nombreux outils et bibliothèques logicielles open source.

Le volet MLOps

La traçabilité totale exigée par les Articles 11 et 12 s'opérationnalise en s'inspirant des bonnes pratiques du monde du développement informatique, par exemple en intégrant des plateformes de gestion du cycle de vie des modèles (ex : *MLflow*) :

- **Versionnage de modèle** : Chaque itération du modèle, ses hyperparamètres, ses configurations d'environnement et les liens vers les identifiants uniques de jeux de données d'entraînement doivent être figés au sein d'un *Model Registry*.
- **Traçabilité des artefacts** : L'enregistrement systématique des courbes d'apprentissage, des matrices de confusion et des rapports de tests d'évaluation *ex ante* garantit la constitution d'un dossier de conformité.

Le volet Data Engineering

L'intégration d'outils de catalogues de données (ex : *OpenMetadata*) permet de centraliser les métadonnées, de documenter précisément la sémantique métier des caractéristiques (*features*) exploitées, d'associer des propriétaires de données aux pipelines et de cartographier graphiquement le lignage des données (*data lineage*) de la source brute jusqu'aux données d'apprentissage et de validation.

Le volet Éthique

Nous avons vu que la question de l'équité traitait du biais d'apprentissage. Cela dit, il s'agit également de performance et de robustesse. En la matière, le recours à des outils spécialisés (ex : *Fairlearn*) est vivement recommandé pour mesurer et atténuer les biais discriminatoires *ex ante* :

- Indicateurs de disparité de traitement sur les attributs sensibles (ex : âge).
- Métriques clés telles que **la parité statistique** (vérifier que le taux de décision positive est égal d'un groupe à l'autre) ou **l'égalité des chances** (vérifier que le taux de vrais positifs est homogène entre les groupes).

Ressources réglementaires

La Commission européenne héberge une « plateforme unique d'information sur la législation sur l'IA (AI Act) » proposant plusieurs services au public :

- **AI Act Explorer** : Outil en ligne conçu pour aider les utilisateurs à parcourir les différents chapitres, annexes et considérants du règlement sur l'IA de manière intuitive.
- **Vérificateur de conformité** : Outil qui aide à évaluer si les systèmes d'IA et les modèles d'IA à usage général satisfont aux exigences fixées par le règlement sur l'IA.
- **Service d'assistance** : Permet aux parties prenantes de poser des questions sur le règlement sur l'IA et de recevoir des réponses d'une équipe d'experts travaillant en étroite coopération avec le Bureau de l'IA.
- La plateforme propose également une **FAQ** et une rubrique publiant les dernières évolutions de la législation.

6 La réglementation des systèmes, est-ce vraiment une nouveauté ?

Nous avons discuté dans les sections précédentes de la façon dont l'AI Act européen s'appuie sur les bonnes pratiques industrielles. Nous proposons maintenant d'analyser la façon dont il s'articule avec les réglementations sectorielles déjà existantes. Bâtie sur les structures, les processus de gouvernance et les méthodologies d'audit sectoriels déjà exigés par les cadres prudentiels (ex : Bâle III, Solvabilité II) ou industriels (ex : Règlement sur les Dispositifs Médicaux ou RDM), la conformité à l'AI Act représente une extension naturelle permettant d'adresser des risques nouveaux. Tour d'horizon non exhaustif.

Le secteur financier se situe au cœur d'une superposition réglementaire dense (ex : Bâle III finalisé, CRR III, CRD VI, IFRS9). Le système de gestion de la qualité exigé pour l'IA à haut risque (ex : scoring de crédit de personnes physiques) ou les exigences de résilience et de sécurité de l'IA font directement écho aux processus de contrôle interne et de gouvernance des risques des différentes normes et directives déjà en vigueur. Côté superviseur, pas de piste d'audit distincts, mais des compléments à celles déjà réalisées par l'ACPR. Idem chez les assureurs, où l'usage de l'IA pour l'évaluation des risques et la tarification dans l'assurance-vie et l'assurance maladie est classé à haut risque. L'évaluation continue exigée par l'AI Act est intégrée au sein du processus d'évaluation interne des risques au titre du régime prudentiel Solvabilité II. Les dispositifs médicaux dotés d'IA nécessitant un audit par un tiers sous la RDM sont également classés à haut risque. Ici non plus, le système de gestion de la qualité exigé par l'AI Act ne nécessite pas d'infrastructure parallèle. Il est conçu pour s'intégrer au système de gestion de la qualité exigé par le RDM. Enfin, l'AI Act s'intègre pleinement dans le système d'homologation en vigueur dans l'industrie automobile. Les obligations relatives à la robustesse technique et à la cybersécurité s'articulent avec les exigences de la norme d'ingénierie de la cybersécurité ISO/SAE 21434. Ainsi, les constructeurs et équipementiers sont invités à coordonner leurs comités internes pour intégrer les nouvelles exigences de l'IA Act au processus classique, évitant ainsi d'avoir à faire certifier le SIA indépendamment de la structure globale du véhicule.

Pour résumer, l'étroite complémentarité entre l'AI Act européen et les exigences sectorielles se matérialise principalement au travers des trois piliers que constituent données, modèles et piste d'audit :

- **Gouvernance et provenance des données** : processus documentés de collecte, de filtrage et de préparation des métadonnées. Si l'Article 10 impose des contrôles stricts sur les données d'entraînement, de validation et de test (ex : représentativité, complétude, détection et correction des biais), cette exigence s'appuie directement sur des cadres de gouvernance des données déjà matures.
- **Gouvernance des Modèles** : validation continue, identification de la dérive des données (*data drift*), représentativité statistique et élimination des biais discriminatoires sont déjà des exigences de la plupart des réglementations sectorielles.
- **Piste d'audit unifiée** : archivage des logs de décision et élaboration du dossier technique complet démontrant le respect des exigences essentielles avant la mise en production ou la mise sur le marché sont les éléments de base d'un dossier d'homologation.

7 Conclusion

Tout au long de cet article, nous avons tenté de démontrer que l'AI Act européen dépasse largement le simple cadre de la contrainte réglementaire pour s'imposer comme un véritable standard de conception et de gouvernance des systèmes automatisés. Après un bref rappel des risques liés à l'utilisation des systèmes d'intelligence artificielle (SIA) et la présentation des différents acteurs du nouveau cadre réglementaire, nous avons adressé notre premier objectif avec une analyse des exigences formulées par le législateur européen au prisme des bonnes pratiques industrielles en matière de SIA, qui encouragent les différentes parties prenantes à adopter des architectures techniques robustes et performantes.

L'intégration des principes de l'approche FAT (*Fairness, Accountability, Transparency*) au cœur du cycle de vie des modèles et systèmes d'IA assure non seulement la protection des droits fondamentaux des utilisateurs, mais garantit également une robustesse opérationnelle accrue face aux risques d'hallucination, de biais discriminatoires ou de cyberattaques. Une attention particulière a été portée au concept d'*Explainable AI* (XAI), visant à dissiper l'opacité inhérente aux modèles complexes qualifiés de « boîtes noires », en mobilisant des méthodes locales et globales évaluées selon des critères stricts de fidélité et de stabilité. Par souci de pédagogie, nous avons accompagné chaque concept de la XIA d'indicateurs statistiques et d'exemples illustratifs.

Conformément à notre deuxième objectif, nous avons montré que l'opérationnalisation de ces directives ne nécessite pas forcément une transformation des pratiques d'ingénierie, car elle repose largement sur des méthodes et des outils déjà adoptés par la plupart des industriels, comme les outils de MLOps, de traçabilité des données et d'évaluation de l'équité algorithmique. De plus, nous avons montré que cette nouvelle législation ne constitue pas une rupture réglementaire isolée, mais s'inscrit plutôt dans le prolongement harmonieux des exigences prudentielles et sectorielles déjà existantes. Que ce soit dans le secteur financier sous l'égide de Bâle III, dans l'assurance avec Solvabilité II, ou encore dans l'industrie médicale et automobile, l'AI Act européen consolide des processus de contrôle interne déjà matures, en évitant les doublons. En effet, la gouvernance rigoureuse des données, la validation continue des modèles algorithmiques et la mise en place d'une piste d'audit unifiée forment désormais un socle commun de conformité intersectorielle.

Enfin, poursuivant notre troisième objectif, nous avons regroupé un certain nombre de ressources techniques (ex : MLOps, data) et méthodologiques (ex : réglementation, éthique) que nous pensons utiles à l'opérationnalisation de l'IA.

Nous pouvons nous attendre à ce que l'application effective de ce texte nécessite des clarifications futures de la part des autorités, notamment concernant la définition d'une « modification substantielle » ou d'un changement de « destination », ou la frontière séparant un fournisseur d'un déployeur. Néanmoins, nous sommes d'avis que ce texte est davantage normatif que prescriptif. L'appropriation proactive de l'AI Act par les concepteurs et utilisateurs ne doit pas être perçue comme un fardeau, mais bien comme un levier stratégique majeur pour bâtir une intelligence artificielle de confiance, soutenable et créatrice de valeur pérenne sur le marché européen.

Remerciements. Cette étude a été financée par Atheïa.
Déclaration d'intérêts. L'auteur déclare n'avoir aucun conflit d'intérêt avec le contenu de cet article.

Références

1. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 1689. European Union.
2. Regulation (EU) 2026/1744 of the European Parliament and of the Council of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI)
3. Guidelines on the definition of an artificial intelligence system established by Regulation (EU) 2024/1689 (AI Act) (2025)
4. Commission Guidelines on the scope of the obligations for providers of general-purpose AI models established by Regulation (EU) 2024/1689 (AI Act) (2025)
5. Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act) (2025)
6. High-Level Expert Group on AI (AI HLEG). (2019).
7. Lundberg, S. M., & Lee, S. I. (2017a). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.
8. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).
9. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., & Viegas, F. (2018, July). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning* (pp. 2668-2677). PMLR.
10. DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., & Wallace, B. C. (2020, July). ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4443-4458).
11. O. Bamigbade, Y. T. Adeshina, K. Kemisola (2024) Ethical and Explainable AI In Data Science for Transparent Decision-Making across Critical Business Operations. International Journal of Engineering Technology Research & Management.
12. H. Zhao et al. (2024) Explainability for Large Language Models: A Survey. ACM Trans. Intell. Syst. Technol. 15, 2, Article 20