Le glossaire de la gouvernance de l'IA (Édition enrichie, 50 termes)

Un instrument gratuit de AI GOVERNANCE CHAIN™. Le vocabulaire est gratuit. Les méthodes appliquées qui l'utilisent constituent le parcours payant.

Abréviations et acronymes

Cette section définit chaque acronyme utilisé dans ce document. Aucune connaissance préalable n'est supposée.

- IA : intelligence artificielle.
- UE : Union européenne.
- ISO/IEC 42001 : la norme internationale de système de management de l'intelligence artificielle, publiée en décembre 2023 par l'Organisation internationale de normalisation (ISO) et la Commission électrotechnique internationale (IEC).
- NIST AI RMF : le cadre de gestion des risques de l'IA (AI Risk Management Framework) publié par le National Institute of Standards and Technology des États-Unis.

Pourquoi ce glossaire existe (le POURQUOI)

Un domaine ne peut pas être gouverné tant qu'il ne peut pas être nommé. Les conseils d'administration, les responsables de la conformité et les ingénieurs parlent régulièrement sans se comprendre, parce qu'ils emploient les mêmes mots pour désigner des choses différentes. Ce glossaire donne à chaque partie prenante un vocabulaire commun, en langage clair, afin que les conversations de gouvernance partent d'un terrain partagé. Il est délibérément écrit pour le non-spécialiste.

Les termes

Gouvernance de l'IA. Le système de politiques, de contrôles, de rôles et de processus de revue qui maintient l'usage de l'IA d'une organisation sûr, licite, responsable et aligné sur ses valeurs.

Responsabilité (accountability). Le principe selon lequel une personne nommée, et non un comité ou un système, répond d'une décision ou d'un résultat donné de l'IA.

IA agentique. Des systèmes d'IA capables d'agir en direction d'un objectif avec une intervention humaine limitée ou nulle, plutôt que de produire uniquement du texte ou des prédictions.

Alignement. Le degré auquel le comportement d'un système d'IA correspond aux intentions et aux valeurs des personnes qui le déploient.

Piste d'audit. Un enregistrement complet et horodaté des décisions, des entrées et des sorties, qui permet de revoir le comportement d'un système d'IA après coup.

Biais. Une distorsion systématique dans les sorties d'un système d'IA qui produit des résultats injustes ou inexacts pour certains groupes ou certains cas.

Capitulation. Un mode de défaillance de sycophancie dans lequel un modèle abandonne une réponse correcte après une objection de l'utilisateur, même lorsque celui-ci a tort.

Dérive de confirmation. Un mode de défaillance de sycophancie dans lequel un modèle déplace progressivement ses réponses vers les préférences exprimées de l'utilisateur plutôt que vers les faits.

Évaluation de la conformité. Un processus formel, exigé par certaines réglementations, visant à démontrer qu'un système d'IA à haut risque satisfait aux exigences définies avant son déploiement.

Contrôle. Une mesure précise, technique ou procédurale, mise en place pour réduire un risque d'IA défini.

Journal de traçabilité des décisions. Un enregistrement structuré expliquant comment et pourquoi une décision influencée par l'IA a été prise, utilisé pour la responsabilité et la revue.

Déploiement. Le moment où un système d'IA passe des tests à un usage réel affectant des personnes ou des opérations réelles.

Dérive. La dégradation progressive des performances d'un système d'IA au fil du temps, à mesure que le monde réel s'éloigne des données sur lesquelles il a été construit.

Explicabilité. La mesure dans laquelle un humain peut comprendre pourquoi un système d'IA a produit une sortie particulière.

Accord fabriqué. Un mode de défaillance de sycophancie dans lequel un modèle invente un soutien ou un consensus en faveur de la position de l'utilisateur alors qu'il n'existe pas.

Biais de flatterie. Un mode de défaillance de sycophancie dans lequel un modèle complimente la contribution de l'utilisateur au lieu de l'évaluer honnêtement.

Modèle de fondation. Un grand modèle d'IA à usage général, entraîné sur des données étendues, qui peut être adapté à de nombreuses tâches spécifiques.

Écart de gouvernance. La distance entre les principes d'IA affichés d'une organisation et ses pratiques opérationnelles réelles.

Garde-fou. Une contrainte qui limite ce qu'un système d'IA est autorisé à faire, conçue pour prévenir les actions nuisibles ou hors du périmètre.

Hallucination. Une sortie d'IA fluide et assurée, mais factuellement fausse ou sans fondement.

Système à haut risque. Au sens du règlement européen sur l'IA, un système d'IA utilisé dans un domaine sensible (comme l'emploi, l'éducation ou les services essentiels) qui emporte des obligations renforcées.

Humain dans la boucle. Une conception dans laquelle un humain examine ou approuve la sortie d'un système d'IA avant qu'elle ne prenne effet.

Humain en supervision. Une conception dans laquelle un humain surveille un système d'IA qui agit de manière autonome et peut intervenir si nécessaire.

ISO/IEC 42001. La première norme internationale de système de management pour l'intelligence artificielle, publiée en décembre 2023, sur laquelle une organisation peut être certifiée.

Incident. Un événement dans lequel un système d'IA cause ou risque de causer un préjudice, une erreur ou une violation d'une politique ou de la loi.

Cycle de vie. La durée complète de la vie d'un système d'IA, de la conception et du développement jusqu'au déploiement, à l'exploitation et à la mise hors service.

Modèle de maturité. Une échelle qui décrit la capacité de gouvernance d'une organisation selon des niveaux définis, de l'absence à l'optimisation.

Fiche de modèle. Un document court décrivant l'usage prévu, les limites et les caractéristiques de performance d'un modèle d'IA.

Surveillance. L'observation continue d'un système d'IA en production afin de détecter la dérive, la défaillance ou un risque émergent.

NIST AI RMF. Le cadre de gestion des risques de l'IA publié par le National Institute of Standards and Technology des États-Unis, organisé autour de quatre fonctions : gouverner, cartographier, mesurer et gérer.

Supervision. Les structures et les rôles par lesquels une organisation encadre son usage de l'IA.

Retour d'expérience. Une revue structurée conduite après un incident pour identifier les causes profondes et prévenir la récurrence.

Pré-déploiement à grande échelle. L'étape juste avant qu'un système d'IA ne soit étendu à un usage large, le moment de plus fort levier pour le gouverner.

Gouvernance des achats. La pratique consistant à appliquer des exigences de gouvernance aux systèmes et outils d'IA qu'une organisation achète plutôt que construit.

Injection de requête. Une attaque dans laquelle une entrée fabriquée amène un système d'IA à ignorer ses instructions ou à se comporter de manière dangereuse.

Test d'intrusion adverse. Le fait de sonder délibérément un système d'IA avec des entrées adverses afin d'en trouver les faiblesses avant qu'elles ne causent un préjudice.

Niveau de risque. Une classification d'un système d'IA selon le niveau de risque qu'il porte, utilisée pour attribuer des obligations proportionnées.

Robustesse. La capacité d'un système d'IA à maintenir des performances fiables dans des conditions inattendues ou adverses.

Échelle du préjudice. Le principe selon lequel une défaillance de l'IA peut affecter de nombreuses personnes à la fois, rendant les petites erreurs lourdes de conséquences à grande échelle.

Jalons de passage. Une méthode de gouvernance dans laquelle un système doit satisfaire à des critères définis à chaque étape du cycle de vie avant de passer à la suivante.

Partie prenante. Toute partie affectée par un système d'IA ou responsable de celui-ci, y compris les utilisateurs, les personnes concernées, les opérateurs et les organes de supervision.

Sycophancie. La tendance d'un système d'IA à dire aux utilisateurs ce qu'ils veulent entendre, en approuvant ou en flattant plutôt qu'en étant exact.

Risque de tiers. Le risque de gouvernance qui naît des systèmes ou des composants d'IA fournis par des fournisseurs externes.

Traçabilité. La capacité de remonter une décision d'IA à travers les données, le modèle et la logique qui l'ont produite.

Transparence. La pratique consistant à rendre l'existence, la finalité et les limites d'un système d'IA claires pour les personnes concernées.

IA digne de confiance. Une IA licite, éthique et robuste, qui se comporte comme prévu tout au long de son cycle de vie.

Cas d'usage. Une application précise d'un système d'IA à une tâche ou à un problème défini.

Validation. Le processus consistant à confirmer qu'un système d'IA fonctionne correctement et en sécurité pour son usage prévu avant le déploiement.

Alignement des valeurs. S'assurer que le comportement d'un système d'IA est cohérent avec les valeurs affichées de l'organisation, et non seulement avec ses objectifs techniques.

Registre des risques fournisseurs. Un enregistrement des fournisseurs d'IA sur lesquels une organisation s'appuie et des risques que chacun introduit.

Où s'arrête la couche gratuite et où commence le parcours payant

Ce glossaire vous donne le QUOI (le vocabulaire commun). La couche payante vous donne le COMMENT : les cadres appliqués, les modèles de notation et les systèmes d'exploitation qui transforment ces termes en un programme de gouvernance opérationnel.

- Pour le COMMENT structuré : le Système d'exploitation de la gouvernance de l'IA.
- Pour l'AVEC-VOUS : The Governance Circle.
- Pour le FAIT-POUR-VOUS : l'Audit de gouvernance pré-déploiement.

Mettez cette page en favori, partagez-la, citez-la. Un langage commun est le point de départ de la gouvernance.

AI GOVERNANCE CHAIN™

Cadre de gouvernance de l'IA

Membre du conseil consultatif de la Harvard Business Review.

Norme documentaire de AI GOVERNANCE CHAIN™ : une gouvernance de l'IA centrée sur l'humain.