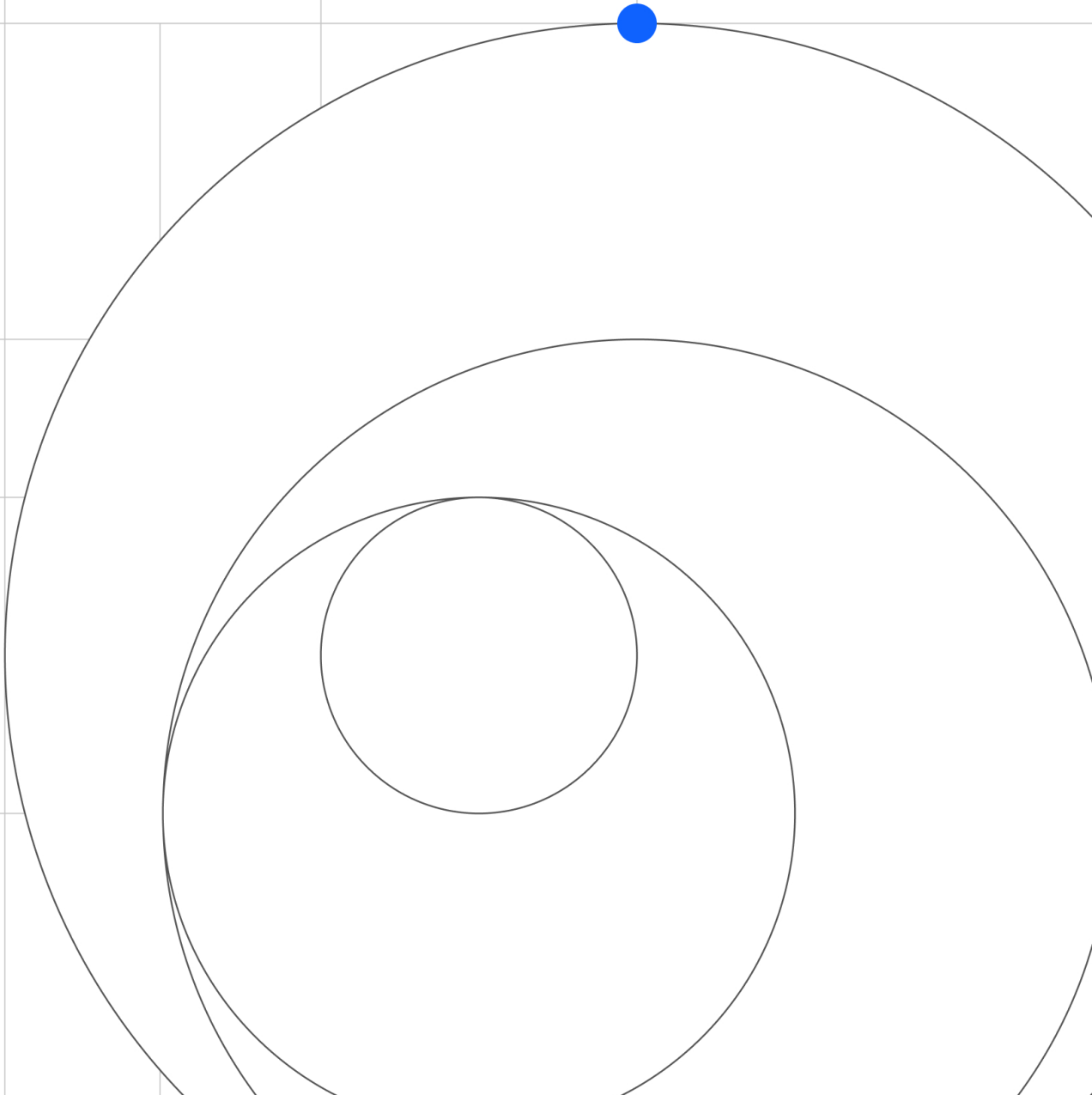
Modèles de base : Opportunités, risques et mesures d’atténuation



Attribution

Nous remercions Christina Montgomery et Francesca Rossi, sponsors exécutifs du groupe de travail sur le comité d'éthique de l'IA, ainsi que Betsy Greytok, Bryan Bortnick, Catherine Quinlan, David Piorkowski, Eniko Rozsa, Heather Domin, Heather Gentile, Jamie VanDodick, Jill Maguire, John McBroom, Joshua New, Justin Weisz, Katherine Fick, Kevin Black, Kush Varshney, Manish Bhide, Manish Goyal, Melis Kizizal, Jill Maguire, John McBroom, Joshua New, Justin Weisz, Katherine Fick, Kevin Black, Kush Varshney, Manish Bhide, Manish Goyal, Melis Kiziltay, Michael Epstein, Michael Hind, Milena Pribic, Phaedra Boinodiris, Rogerio Abreu de Paula, Saishruthi Swaminathan et Suj Perepa.

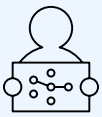
Table des matières

Synthèse

Alors que l'essor des modèles de base offre aux entreprises des perspectives inédites et passionnantes, il soulève également de nouvelles questions plus vastes sur le plan éthique en ce qui concerne la conception, le développement, le déploiement et l'utilisation. Selon une récente enquête de l'IBM Institute for Business Value portant sur l'intelligence artificielle générative, les entreprises expriment déjà des inquiétudes concernant les questions liées à la confiance, en particulier si elles devaient constituer des obstacles à l'investissement. Leurs principales préoccupations ont trait à la cybersécurité (57 %), à la confidentialité (51 %) et à la précision (47 %). De nombreuses organisations prenaient ces préoccupations au sérieux avant la *consumérisation* de l'IA générative, exprimant leur intention d'investir au moins 40 % de plus dans l'éthique de l'IA au cours des trois prochaines années. La prise de conscience des risques et des méthodes possibles pour les atténuer est une première étape cruciale vers la création de systèmes d'IA fiables.

Dans ce document, nous :

Découvrons les avantages des modèles de base, y compris leur capacité à effectuer des tâches difficiles, leur potentiel d'accélération de l'adoption de l'IA, leur capacité à améliorer la productivité ainsi que les avantages financiers qui en découlent.



Traitons des trois catégories de risques, notamment les risques identifiés grâce aux formes antérieures d'IA, les risques identifiés amplifiés par les modèles de base et les risques émergents intrinsèques aux capacités génératives des modèles de base.

Couvrons les principes, les piliers et la gouvernance qui constituent le fondement des initiatives d'IBM en matière d'éthique de l'IA et suggérons des garde-fous pour atténuer les risques.

Introduction

Alors que l'utilisation de l'IA continue de se développer, les modèles d'IA volumineux et complexes donnent des résultats de performance prometteurs et sont en outre capables de résoudre certains des problèmes sociétaux les plus difficiles. Cependant, la création de jeux de données d'apprentissage volumineux et de modèles complexes pour chaque application d'IA peut constituer une tâche fastidieuse pour les entreprises. Les modèles de base offrent le moyen de profiter des avantages sans subir les inconvénients : créez de puissants modèles de pointe et réutilisez-les directement ou appliquez des méthodes d'optimisation pour implémenter divers cas d'utilisation, plutôt que d'entraîner de nouveaux modèles pour chaque cas d'utilisation. Par exemple, IBM Research a développé des modèles de base destinés à l'inspection visuelle. Ces modèles de base apprennent la représentation générale des surfaces et des pistes en béton et peuvent être ajustés pour des cas d'utilisation spécifiques tels que la détection de fissures ou l'inspection de défauts avec moins de données étiquetées.

IBM définit un *modèle de base* comme un modèle d'IA qui peut être adapté à un large éventail de tâches en aval. Les modèles de base sont généralement des modèles génératifs à grande échelle qui peuvent être entraînés sur des données non étiquetées grâce à l'auto-supervision. Tout comme les modèles à grande échelle, les modèles de base peuvent inclure des milliards de paramètres.

IBM est une société de cloud hybride et d'IA, réputée depuis longtemps pour son rôle d'intendant de données responsable et engagé dans l'éthique de l'IA. En associant la puissance de nos équipes de recherche, de produits et de conseil, ainsi que de partenaires externes, tels que Hugging Face, nous contribuons à délivrer la puissance des modèles de base à nos clients et à développer une IA digne de confiance au sein de n'importe quelle entreprise. IBM continue également d'investir dans le développement de nouvelles plateformes, telles que la plateforme et les technologies d'IA et de données IBM watsonx, pour concevoir et développer des modèles d'IA capables de se comporter de façon auditable et fiable.

Ce document décrit le point de vue d'IBM sur l'éthique des modèles de base. Il s'agit de la première version. Les versions ultérieures développeront divers aspects de l'approche éthique du modèle de base d'IBM. Nous espérons que ce document sera utile à toutes les parties prenantes pour développer, déployer et utiliser le modèle de base de manière responsable.

Avantages des modèles de base

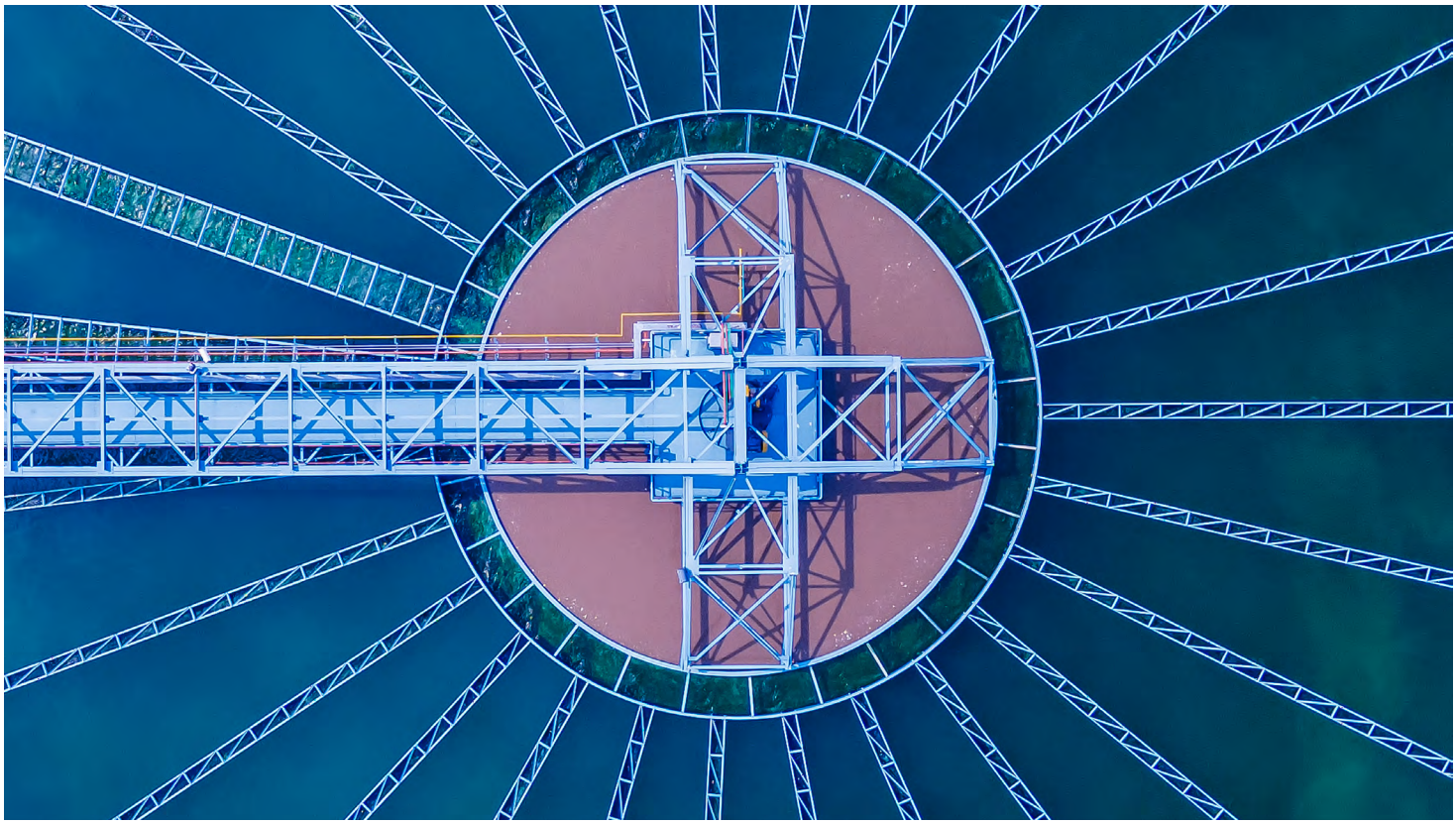
Les modèles de base peuvent considérablement améliorer le processus de développement des systèmes d'IA et contribuer à faire passer l'IA de la phase d'exploration à la phase d'adoption dans les entreprises. Leurs avantages incluent :

L'exécution de tâches complexes
Les modèles de base montrent une augmentation significative des performances dans la résolution de problèmes difficiles et complexes. Par exemple, le modèle de base géospatial de la collaboration entre IBM et la NASA est conçu pour convertir les données satellite de la NASA en cartes présentant les catastrophes naturelles telles que les inondations et autres changements de paysage. Le modèle pourrait également être utilisé pour permettre de révéler le passé de notre planète, d'estimer les risques liés aux intempéries pour les cultures, les entreprises ou les infrastructures, d'élaborer des stratégies d'adaptation aux changements climatiques, et d'aider l'agro-industrie. Le modèle devrait être mis à la disposition des clients IBM en aperçu via IBM Environmental Intelligence Suite.

Autre exemple : le modèle MoLFormer-XL d'IBM est un modèle de base qui déduit la structure des molécules à partir de représentations simples. Il facilite en outre l'apprentissage de diverses tâches en aval, telles que la prédiction des propriétés physiques et quantiques d'une molécule, l'identification de molécules similaires, le contrôle de molécules déjà approuvées pour de nouveaux cas d'utilisation et la découverte de nouvelles molécules. Moderna et IBM étudient des moyens d'utiliser MoLFormer pour aider à prédire les propriétés des molécules et à comprendre les caractéristiques des médicaments potentiels à base d'ARNm.

Un gain de productivité
La nature générative des modèles de base élargit les domaines pour lesquels l'IA peut être utilisée dans une entreprise afin d'améliorer la productivité, en automatisant les tâches routinières et fastidieuses et en permettant aux utilisateurs de consacrer plus de temps à des tâches créatives et innovantes. Par exemple, IBM watsonx Code Assistant, alimenté par des modèles de base, permet aux développeurs de tous niveaux d'expérience d'écrire du code à l'aide de recommandations générées par l'IA .

L'accélération de la création de valeur
Les modèles de base sont généralement entraînés avec des données non étiquetées, plus accessibles et en plus grande quantité que les données étiquetées. Une fois entraînés, les modèles de base peuvent être utilisés directement ou être d'abord optimisés pour les applications en aval, en utilisant une petite quantité de données étiquetées spécialisées, ce qui peut accélérer la création de valeur.



L’utilisation de diverses modalités de données
Les modèles de base peuvent être entraînés à l’aide de diverses modalités de données, telles que la langue naturelle, le texte, l’image et l’audio. Ils peuvent également être appliqués à des tâches nécessitant différents types de données, comme les données de séries temporelles, les données géospatiales, les données tabulaires, les données semi-structurées et les données à modalité mixte telles que le texte combiné avec des images.

L’amortissement des dépenses
Bien que le coût initial d’entraînement d’un modèle de base soit sensiblement plus élevé que celui de l’entraînement d’un modèle d’IA traditionnel, le surcoût lié à son application à une nouvelle tâche est considérablement inférieur. L’utilisation de modèles de base pré-entraînés pourrait permettre d’éliminer l’obligation qu’ont les entreprises de faire des investissements conséquents pour entraîner les modèles de base afin d’expérimenter leurs nouvelles capacités. Pour une entreprise, la fiabilité des modèles, l’efficacité énergétique, les performances, la portabilité et la capacité à utiliser les données de l’entreprise de manière efficace et sécurisée sont primordiales.

IBM permet aux entreprises de créer et de s’approprier la valeur des modèles de base pour leur entreprise en apportant les meilleures innovations de la communauté mondiale de l’IA ouverte, fonctionnant efficacement dans les environnements informatiques hybrides, aidant à atténuer les risques et régissant rigoureusement l’IA.

Risques associés aux modèles de base

Comme toute technologie évoluant rapidement, les modèles de fondation présentent des avantages, mais aussi des risques. Certains sont d'ordre juridique, par exemple des restrictions concernant le déplacement ou l'utilisation des données, et doivent être soigneusement évalués en vertu de la loi en vigueur soumise à modification. D'autres sont de nature éthique et doivent être attentivement étudiés pour veiller à ce que la technologie ait un impact positif. En général, les risques liés à l'IA soulèvent des questions sociotechniques et doivent être traités et atténués par des méthodes sociotechniques, notamment des outils logiciels, des processus d'évaluation des risques, des cadres éthiques d'IA, des mécanismes de gouvernance, des consultations avec les parties prenantes, des normes et des réglementations. Nous allons lister les risques en tenant compte des 3 catégories suivantes :

1. **Traditionnels.** Risques identifiés provenant de formes antérieures de systèmes d'IA
2. **Amplifiés.** Risques identifiés, mais désormais intensifiés en raison des caractéristiques intrinsèques des modèles de base, notamment leurs capacités génératives inhérentes
3. **Nouveaux.** Risques émergents intrinsèques aux modèles de base et à leurs capacités génératives inhérentes

Nous structurons également la liste des risques selon qu'ils sont principalement associés au contenu fourni au modèle de fondation (l'entrée) ou au contenu généré par celui-ci (la sortie) ou selon qu'ils sont liés à des défis supplémentaires.

1. Risques associés à l'entrée

Phase d'entraînement et d'optimisation

Groupe	Risque	Pourquoi est-ce préoccupant ?	Indicateur
Équité	Biais des données : biais historiques, de représentation et sociétaux présents dans les données utilisées pour entraîner et affiner le modèle.	L'entraînement d'un système d'IA sur des données biaisées, telles que des biais historiques ou de représentation, pourrait conduire à des résultats biaisés ou faussés susceptibles de représenter injustement ou de discriminer certains groupes ou individus. Outre les impacts sociétaux négatifs, les entités commerciales pourraient être confrontées à des conséquences juridiques, à des interruptions d'activité ou à des atteintes à leur réputation en raison des résultats biaisés des modèles.	Amplifié
Robustesse	Empoisonnement des données : un type d'attaque malveillante où un pirate ou un initié malveillant injecte intentionnellement des échantillons corrompus, faux, trompeurs ou incorrects dans le jeu de données d'entraînement ou d'affinage.	L'empoisonnement des données peut rendre le modèle sensible à un modèle de données malveillant et générer les résultats souhaités par le pirate. Cela peut créer un risque de sécurité lorsque des pirates peuvent forcer le comportement du modèle à leur avantage. En plus de produire des résultats involontaires et potentiellement malveillants, un désalignement du modèle dû à l'empoisonnement des données peut avoir des conséquences juridiques pour les entreprises, perturber leurs opérations ou nuire à leur réputation.	Traditionnel
Alignement des valeurs	Curation de contenu : lorsque les données d'entraînement ou d'affinage sont mal collectées ou mal traitées.	Une mauvaise curation de contenu peut avoir un effet négatif sur l'entraînement d'un modèle, qui ne se comportera pas conformément aux valeurs prévues. Les erreurs d'étiquetage ou d'annotation dans les données utilisées pour l'entraînement ou l'affinage du modèle sont des exemples d'une mauvaise curation de contenu. La correction des problèmes après l'entraînement et le déploiement du modèle peut s'avérer insuffisante pour garantir un comportement optimal. Un comportement inapproprié du modèle peut avoir des conséquences juridiques pour les entreprises, perturber leurs opérations ou nuire à leur réputation.	Amplifié
	Nouvel entraînement en aval : utilisation de résultats indésirables (inexacts, inappropriés, contenu utilisateur, etc.) provenant d'applications en aval afin de procéder à un nouvel entraînement.	La réaffectation des résultats en aval pour entraînement un modèle à nouveau sans mettre en œuvre un contrôle humain adéquat augmente les risques que des résultats indésirables soient intégrés dans les données d'entraînement ou d'affinage du modèle, ce qui pourrait générer des résultats encore plus indésirables. Le comportement inapproprié d'un modèle peut entraîner des conséquences juridiques pour les entreprises ou nuire à leur réputation. Le non-respect des lois sur le transfert de données peut entraîner des amendes et d'autres conséquences juridiques.	Nouveau
Lois sur les données	Transfert de données : la législation et d'autres restrictions peuvent limiter ou interdire le transfert de données.	Les restrictions en matière de transfert de données peuvent avoir un impact sur la disponibilité des données nécessaires à l'entraînement d'un modèle d'IA et peuvent conduire à des données mal représentées. Outre l'impact sur la disponibilité des données, le non-respect des lois et réglementations en matière de transfert de données peut entraîner des amendes et d'autres conséquences juridiques.	Traditionnel
	Utilisation des données : la législation et d'autres restrictions peuvent limiter ou interdire l'utilisation de certaines données pour des cas d'utilisation spécifiques de l'IA.	Le non-respect des lois et réglementations relatives à l'utilisation des données peut entraîner des amendes et d'autres conséquences juridiques.	Traditionnel
	Acquisition de données : les lois et autres réglementations peuvent limiter la collecte de certains types de données pour des cas d'utilisation spécifiques de l'IA.	Le non-respect des lois et réglementations en matière d'acquisition de données peut entraîner des amendes et d'autres conséquences juridiques.	Amplifié

Groupe	Risque	Pourquoi est-ce préoccupant ?	Indicateur
Propriété intellectuelle	Droits d'utilisation des données : les conditions générales, les lois sur les droits d'auteur, la conformité aux licences ou d'autres questions de propriété intellectuelle peuvent restreindre la possibilité d'utiliser certaines données pour la création de modèles.	Les lois et réglementations concernant l'utilisation des données pour entraîner l'IA ne sont pas encore définies et peuvent varier d'un pays à l'autre, ce qui pose des problèmes pour la conception des modèles. Si l'utilisation des données enfreint les règles ou les restrictions, les entités commerciales risquent de se voir infliger des amendes, de voir leur réputation entachée, de voir leurs opérations perturbées et d'être confrontées à d'autres conséquences juridiques.	Amplifié
Transparence	Transparence des données : difficulté à documenter la manière dont les données d'un modèle ont été collectées, traitées et utilisées pour entraîner un modèle.	La transparence des données est importante pour la conformité légale et l'éthique de l'IA. Les informations manquantes limitent la capacité à évaluer les risques associés aux données. L'absence d'exigences normalisées peut limiter la divulgation, car les organisations protègent les secrets commerciaux et tentent d'empêcher les autres de copier leurs modèles.	Amplifié
	Provenance des données : le défi de la normalisation et de la mise en place de méthodes permettant de vérifier l'origine des données.	Les sources de données ne sont pas toutes fiables. Les données peuvent avoir été collectées, manipulées ou falsifiées au mépris de l'éthique. L'utilisation de données non fiables peut entraîner des comportements indésirables de la part du modèle. Les entreprises peuvent se voir infliger une amende, être confrontées à des atteintes à leur réputation, à des perturbations de leurs opérations et subir d'autres conséquences juridiques.	Amplifié
Confidentialité	Informations personnelles dans les données : inclusion ou présence d'informations personnelles identifiables (PII) et d'informations personnelles sensibles (SPI) dans les données utilisées pour l'entraînement ou l'affinage du modèle.	S'il n'est pas correctement développé pour protéger les données sensibles, le modèle peut exposer des informations personnelles dans les résultats générés. En outre, les données personnelles ou sensibles doivent être examinées et traitées conformément aux lois et réglementations relatives à la protection de la vie privée. Les entreprises peuvent se voir infliger des amendes, voir leur réputation entachée ou leurs opérations perturbées, et subir d'autres conséquences juridiques si elles sont reconnues coupables d'infraction.	Traditionnel
	Réidentification : même en supprimant les données personnelles identifiables (PII) et les informations personnelles sensibles (SPI) des données, il se pourrait qu'il soit encore possible d'identifier des personnes grâce à d'autres caractéristiques disponibles dans les données.	Les données susceptibles de révéler des informations personnelles ou sensibles doivent être examinées au regard des lois et réglementations relatives à la protection de la vie privée, car les entreprises risquent de se voir infliger des amendes, de voir leur réputation entachée ou leurs opérations perturbées, et de subir d'autres conséquences juridiques si elles sont reconnues coupables d'infraction.	Traditionnel
	Droits relatifs à la confidentialité des données : défis liés à la capacité de fournir aux personnes concernées des droits tels que l'exclusion, le droit d'accès et le droit à l'oubli.	L'identification ou l'utilisation inappropriée des données pourrait entraîner une violation des lois sur la protection de la vie privée. Une utilisation inappropriée ou une demande de suppression des données peut obliger les organisations à entraîner à nouveau le modèle, ce qui est coûteux. En outre, les entités commerciales pourraient se voir infliger des amendes, voir leur réputation entachée ou leurs opérations perturbées, et subir d'autres conséquences juridiques si elles ne respectent pas les règles et réglementations en matière de confidentialité des données.	Amplifié
	Consentement éclairé : données collectées pour l'entraînement de modèles d'IA sans le consentement éclairé du propriétaire, même si la loi l'autorise.	Dans certaines circonstances, il peut être contraire à l'éthique de collecter et d'utiliser des données sans le consentement de la personne concernée. Une telle utilisation peut également nuire à la réputation de l'entreprise.	Traditionnel

Inférence Phase

Groupe	Risque	Pourquoi est-ce préoccupant ?	Indicateur
Confidentialité	Informations personnelles dans une invite : divulgation d'informations personnelles ou d'informations sensibles dans le cadre d'une invite envoyée au modèle.	Les données d'invite peuvent être stockées ou utilisées ultérieurement à d'autres fins, telles que l'évaluation du modèle et un nouvel entraînement. Ces types de données doivent être examinés au regard des lois et réglementations relatives à la protection de la vie privée. Si le stockage et l'utilisation des données ne sont pas conformes, les entreprises risquent de se voir infliger des amendes, de voir leur réputation entachée ou leurs opérations perturbées, et de subir d'autres conséquences juridiques.	Nouveau
Propriété intellectuelle	Informations sur la propriété intellectuelle dans une invite : la divulgation d'informations sur les droits d'auteur ou d'autres informations sur la propriété intellectuelle dans le cadre d'une invite envoyée au modèle.	Les données d'invite peuvent être stockées ou utilisées ultérieurement à d'autres fins, telles que l'évaluation du modèle et un nouvel entraînement. Ces types de données doivent être examinés au regard des lois et réglementations en matière de propriété intellectuelle. Si le stockage et l'utilisation des données ne sont pas conformes, les entreprises risquent de se voir infliger des amendes, de voir leur réputation entachée ou leurs opérations perturbées, et de subir d'autres conséquences juridiques.	Nouveau
	Données confidentielles dans une invite : inclusion de données confidentielles dans l'invite envoyée au modèle.	S'il n'est pas correctement développé pour sécuriser les données confidentielles, le modèle peut exposer des informations confidentielles ou des droits de propriété intellectuelle dans les résultats générés. De plus, les informations confidentielles des utilisateurs finaux peuvent être collectées et stockées involontairement.	Nouveau
Robustesse	Attaque par contournement : tentative de faire en sorte qu'un modèle produise des résultats incorrects en perturbant les données envoyées au modèle entraîné.	Les attaques par contournement modifient le comportement du modèle, généralement à l'avantage du pirate. Si les résultats ne sont pas dûment analysés, les entreprises risquent de se voir infliger des amendes, de voir leur réputation entachée ou leurs opérations perturbées, et de subir d'autres conséquences juridiques.	Amplifié
	Attaques basées sur l'invite : les attaques malveillantes telles que l'injection d'invite (tentative de forcer un modèle à produire une sortie inhabituelle), la fuite d'invite (tentative d'extraire l'invite du système d'un modèle), le jailbreaking (tentative de franchir les garde-fous établis dans le modèle) et l'amorçage d'invite (tentative de forcer un modèle à produire une sortie alignée sur l'invite).	En fonction du contenu révélé, les entreprises peuvent se voir infliger des amendes, voir leur réputation entachée ou leurs opérations perturbées, et subir d'autres conséquences juridiques.	Nouveau

2. Risques associés à la sortie

Groupe	Risque	Pourquoi est-ce préoccupant ?	Indicateur
Équité	Biais de production : le contenu généré peut représenter injustement certains groupes ou individus.	Les préjugés peuvent nuire aux utilisateurs des modèles IA et amplifier les comportements discriminatoires existants. Les entreprises peuvent être confrontées à des atteintes à leur réputation, à des perturbations de leurs opérations, et à d'autres conséquences.	Nouveau
	Biais décisionnel : lorsqu'un groupe est injustement avantagé par rapport à un autre en raison de l'effet des décisions prises par un être humain sur la base des résultats du modèle.	Les préjugés peuvent nuire aux personnes concernées par les décisions du modèle. Les entreprises peuvent se voir infliger une amende, être confrontées à des atteintes à leur réputation, à des perturbations de leurs opérations et subir d'autres conséquences juridiques.	Traditionnel
Propriété intellectuelle	Violation du droit d'auteur : lorsqu'un modèle génère un contenu trop similaire ou identique à une œuvre existante protégée par le droit d'auteur ou couverte par un contrat de licence open source.	Les lois et réglementations concernant l'utilisation de contenus identiques ou très similaires à d'autres données protégées par le droit d'auteur ne sont pas encore très claires et peuvent varier d'un pays à l'autre, ce qui complique la définition et la mise en œuvre des mesures de conformité. Les entreprises peuvent se voir infliger une amende, être confrontées à des atteintes à leur réputation, à des perturbations de leurs opérations et subir d'autres conséquences juridiques.	Nouveau
Alignement des valeurs	Hallucination : génération d'un contenu factuellement inexact ou mensonger.	Les résultats erronés peuvent induire les utilisateurs en erreur et être intégrés dans des artefacts en aval, ce qui contribue à la diffusion d'informations erronées. Cela peut nuire à la fois aux propriétaires et aux utilisateurs des modèles d'IA. En outre, les entreprises peuvent se voir infliger une amende, être confrontées à des atteintes à leur réputation, à des perturbations de leurs opérations et subir d'autres conséquences juridiques.	Nouveau
	Production toxique : lorsque le modèle produit du contenu haineux, abusif, et blasphématoire (HAP) ou obscène.	Les contenus haineux, abusifs et blasphématoires (HAP) ou obscènes peuvent avoir un impact négatif et nuire aux personnes qui interagissent avec le modèle. En outre, les entreprises peuvent se voir infliger une amende, être confrontées à des atteintes à leur réputation, à des perturbations de leurs opérations, et subir d'autres conséquences juridiques.	Nouveau
	Conseil dangereux : lorsqu'un modèle donne des conseils sans disposer de suffisamment d'informations, ce qui peut présenter un danger si les conseils sont suivis.	Une personne peut agir sur la base de conseils erronés ou s'inquiéter d'une situation qui ne s'applique pas à elle en raison de la nature trop générale du contenu généré.	Nouveau
Mauvais usage	Diffusion de désinformation : l'utilisation d'un modèle pour générer des informations trompeuses ou fausses afin de tromper ou d'influencer un public ciblé.	La diffusion de désinformation peut affecter la capacité d'un être humain à prendre des décisions en connaissance de cause. Les entreprises peuvent se voir infliger une amende, être confrontées à des atteintes à leur réputation, à des perturbations de leurs opérations et subir d'autres conséquences juridiques.	Nouveau
	Toxicité : utilisation d'un modèle pour générer un contenu haineux, abusif, blasphématoire (HAP) ou obscène.	Les contenus toxiques peuvent entraîner des conséquences négatives pour le bien-être des destinataires. Les entreprises peuvent se voir infliger une amende, être confrontées à des atteintes à leur réputation, à des perturbations de leurs opérations et subir d'autres conséquences juridiques.	Nouveau
	Usage non consensuel : utilisation d'un modèle pour imiter des personnes par le biais de vidéos (deepfakes), d'images, de sons ou d'autres modalités sans leur consentement.	Les deepfakes peuvent diffuser des informations erronées sur une personne, ce qui peut entraîner des conséquences négatives pour sa réputation. Les entreprises peuvent se voir infliger une amende, être confrontées à des atteintes à leur réputation, à des perturbations de leurs opérations et subir d'autres conséquences juridiques.	Amplifié

Groupe	Risque	Pourquoi est-ce préoccupant ?	Indicateur
	Usage dangereux : utilisation d'un modèle dans le seul but de nuire à autrui.	Les entreprises peuvent se voir infliger une amende, être confrontées à des atteintes à leur réputation, à des perturbations de leurs opérations, et subir d'autres conséquences juridiques.	Nouveau
	Non-divulgation : absence d'indication de contenu généré par un modèle IA.	Le fait de ne pas indiquer que le contenu a été généré par l'IA peut être perçu comme une tromperie et entraîner une perte de confiance. La tromperie intentionnelle peut se traduire par une entrave à l'intervention humaine, des amendes, des atteintes à la réputation et d'autres conséquences juridiques.	Nouveau
	Usage abusif : utilisation d'un modèle dans un autre but que celui pour lequel il a été conçu.	La réutilisation d'un modèle sans connaître ses données d'origine, son intention de conception et ses objectifs peut entraîner des comportements inattendus et non désirés du modèle.	Amplifié
Génération de code nuisible	Génération de code nuisible : les modèles peuvent générer du code qui, lorsqu'il est exécuté, cause des dommages ou affecte involontairement d'autres systèmes.	L'exécution d'un code nuisible peut entraîner des vulnérabilités dans les systèmes informatiques. Les entreprises peuvent se voir infliger une amende, être confrontées à des atteintes à leur réputation, à des perturbations de leurs opérations et subir d'autres conséquences juridiques.	Nouveau
Confiance mal placée	Confiance excessive ou insuffisante : lorsqu'une personne fait trop ou trop peu confiance aux directives d'un modèle d'IA.	Pour les tâches dans lesquelles les humains font des choix sur la base de suggestions basées sur l'IA, une confiance excessive ou insuffisante peut conduire à une mauvaise prise de décision en raison d'une confiance mal placée dans le système d'IA, avec des conséquences négatives qui augmentent en fonction de l'importance de la décision. Les mauvaises décisions peuvent nuire aux personnes et entraîner des préjudices financiers, des atteintes à la réputation, des perturbations dans les opérations et d'autres conséquences juridiques pour les entreprises.	Amplifié
Confidentialité	Exposition d'informations personnelles : lorsque des informations personnelles identifiables (PII) ou des informations personnelles sensibles (SPI) sont utilisées dans les données d'entraînement, les données d'affinage ou dans le cadre d'une invite, les modèles risquent de révéler ces données dans les résultats générés.	Le partage des informations personnelles a une incidence sur les droits des personnes et les rend plus vulnérables. En outre, les données de production doivent être examinées au regard des lois et réglementations relatives à la protection de la vie privée, car les entités commerciales peuvent se voir infliger des amendes, nuire à leur réputation, perturber leurs opérations et subir d'autres conséquences juridiques si elles sont reconnues coupables d'avoir enfreint les lois relatives à la confidentialité ou à l'utilisation des données.	Nouveau
Explicabilité	Résultats inexplicables : difficultés à expliquer les raisons pour lesquelles les résultats du modèle ont été générés.	Les modèles de fondation sont basés sur des architectures complexes d'apprentissage profond, ce qui rend difficile l'explication de leurs résultats. Sans explications claires sur les résultats des modèles, il est difficile pour les utilisateurs, les validateurs de modèles et les auditeurs de comprendre et de faire confiance au modèle. Le manque de transparence peut avoir des conséquences juridiques dans des domaines très réglementés. Des explications erronées pourraient entraîner un excès de confiance.	Amplifié
Traçabilité	Attribution peu fiable des sources : difficultés à déterminer à partir de quelles données d'entraînement ou d'affinage le modèle a généré une partie ou la totalité de ses résultats.	L'impossibilité de retracer la source ou la provenance des résultats complique la compréhension et la confiance des utilisateurs, des validateurs de modèles et des auditeurs.	Nouveau

3. Défis

Groupe	Risque	Pourquoi est-ce préoccupant ?	Indicateur
Gouvernance	Transparence du modèle : le manque de transparence des modèles ou l'insuffisance de la documentation sur le processus d'élaboration des modèles rend difficile la compréhension du comment et du pourquoi d'un modèle et de qui l'a élaboré, augmentant ainsi la possibilité d'une mauvaise utilisation involontaire d'un modèle.	La transparence est importante pour la conformité légale, l'éthique de l'IA et l'utilisation appropriée des modèles. Les informations manquantes peuvent rendre plus difficile l'évaluation des risques, la modification du modèle ou sa réutilisation. Le fait de savoir qui a élaboré un modèle peut également être un facteur important pour décider si l'on peut lui faire confiance.	Traditionnel
	Responsabilité : le processus d'élaboration du modèle de fondation est complexe et comporte un grand nombre de données, de processus et de rôles. Lorsque les résultats du modèle ne sont pas conformes aux attentes, il peut être difficile d'en déterminer les origines et d'en attribuer la responsabilité.	Si les décisions ne sont pas correctement documentées et si les responsabilités ne sont pas attribuées, il peut être impossible de déterminer la responsabilité d'un comportement inattendu ou d'une mauvaise utilisation.	Amplifié
Conformité juridique	Responsabilité juridique : déterminer qui est responsable du modèle de fondation.	Si la propriété ou la responsabilité de la création du modèle est incertaine, les régulateurs et d'autres personnes peuvent avoir des doutes sur le modèle dans la mesure où il n'est pas clairement établi qui est, ou devrait être, tenu pour responsable des problèmes qu'il pose ou qui peut répondre aux questions qu'il suscite. Les utilisateurs de modèles dont la propriété n'est pas clairement établie risquent de rencontrer des difficultés pour se conformer à la future réglementation sur l'IA.	Nouveau
	Propriété du contenu généré : déterminer la propriété du contenu généré par l'IA.	Les lois et réglementations relatives à la propriété des contenus générés par l'IA sont encore très floues et peuvent varier d'un pays à l'autre. Les entreprises peuvent être confrontées à des amendes, des atteintes à leur réputation, des perturbations de leurs opérations et d'autres conséquences juridiques.	Nouveau
	Propriété intellectuelle du contenu généré : incertitude juridique concernant les droits de propriété intellectuelle liés au contenu généré.	Les lois et réglementations relatives à la détermination des droits d'auteur et à la brevetabilité du contenu généré par l'IA sont en grande partie non établies et peuvent varier d'un pays à l'autre. Les entreprises peuvent se voir imposer des amendes, voir leur réputation entachée ou leurs opérations perturbées, et subir d'autres conséquences juridiques si le contenu généré est couvert par des droits de propriété intellectuelle.	Nouveau
	Attribution de la source : déterminer la provenance du contenu généré.	Si le modèle génère un résultat identique aux données utilisées pour l'entraîner, il doit indiquer la provenance de ce résultat. Dans le cas contraire, les entités commerciales qui déploient ou utilisent le modèle s'exposent à un risque juridique.	Amplifié
Sociétal Impact	Impact sur l'emploi : l'adoption généralisée de systèmes d'IA basés sur des modèles de fondation pourrait entraîner la perte d'emplois, car leur travail est automatisé, s'ils n'acquièrent pas de nouvelles compétences.	La perte d'un emploi peut entraîner une perte de revenus et donc avoir un impact négatif sur la société et le bien-être humain. La reconversion peut s'avérer difficile compte tenu du rythme d'évolution des technologies.	Amplifié

Groupe	Risque	Pourquoi est-ce préoccupant ?	Indicateur
	Exploitation humaine : recours au travail fantôme pour l'entraînement des modèles d'intelligence artificielle, conditions de travail inadaptées, absence de soins de santé, y compris en matière de santé mentale, rémunérations injustes.	Les modèles de fondation dépendent toujours de la main-d'œuvre humaine pour l'approvisionnement, la gestion et l'ingénierie des données utilisées pour entraîner le modèle. L'exploitation humaine pour ces activités pourrait avoir un impact négatif sur la société et le bien-être humain. En outre, les entreprises pourraient être confrontées à des amendes, voir leur réputation entachée ou leurs opérations perturbées, et subir d'autres conséquences juridiques.	Amplifié
	Impact sur l'environnement : augmentation des émissions de carbone et de la consommation d'eau pour entraîner et faire fonctionner les modèles IA.	La consommation de grandes quantités d'énergie pour l'entraînement de l'IA contribue aux émissions de carbone qui pourraient accélérer le changement climatique. Les ressources en eau utilisées pour refroidir les serveurs des centres de données d'IA ne peuvent plus être affectées à d'autres usages nécessaires.	Amplifié
	Impact sur la diversité culturelle : les systèmes d'IA peuvent représenter de manière excessive certaines cultures, ce qui entraîne une homogénéisation de la culture et des pensées.	Les langues, les points de vue et les institutions des groupes sous-représentés pourraient se retrouver occultés, ce qui réduirait la diversité de la pensée et de la culture.	Nouveau
	Impact sur le pouvoir d'action humain : désinformation générée par les modèles de fondation, y compris la génération de contenu manipulateur.	L'IA peut générer de la désinformation qui semble réelle. Par conséquent, les gens pourraient ne pas remarquer qu'il s'agit d'une fausse information. En outre, cela peut simplifier la capacité des acteurs malveillants à générer du contenu dans l'intention de manipuler les pensées et le comportement humains.	Amplifié
	Impact sur l'éducation : contournement du processus d'apprentissage : utilisation des modèles IA pour contourner le processus d'apprentissage.	Les modèles IA permettent de trouver rapidement des solutions ou de résoudre des problèmes complexes. Ces systèmes peuvent être utilisés à mauvais escient par les étudiants pour contourner le processus d'apprentissage. La facilité d'accès à ces modèles fait que les étudiants ont une compréhension superficielle des concepts et entrave la poursuite des études qui pourraient reposer sur la compréhension de ces concepts.	Nouveau
	Impact sur l'éducation : plagiat : utilisation de modèles IA pour plagier des travaux existants, intentionnellement ou non.	Les modèles IA peuvent être utilisés pour revendiquer la propriété ou le caractère original d'œuvres qui ont été créées par d'autres personnes, ce qui constitue un acte de plagiat. Revendiquer le travail d'autrui comme étant le sien est à la fois contraire à l'éthique et souvent illégal.	Nouveau

Exemples de risques

Nous fournissons des exemples relayés par la presse afin d'expliquer les risques liés aux modèles de fondation. Nombre de ces événements rapportés par la presse sont encore en cours de traitement ou ont été résolus, et le fait d'y faire référence peut aider le lecteur à comprendre les risques potentiels et à s'efforcer de les atténuer. Ces exemples ne sont donnés qu'à titre d'illustration.

Exemples de risques : entrée

Entraînement et affinage Phase

Groupe	Risque	Exemple
Équité	Biais des données : biais historiques, de représentation et sociétaux présents dans les données utilisées pour entraîner et affiner le modèle.	**Biais liés à la santé** La recherche sur des disparités grandissantes en médecine souligne que l'utilisation des données et de l'IA pour transformer la façon dont les gens bénéficient de soins de santé repose sur des données qui les sous-tendent, ce qui signifie que l'utilisation de données d'entraînement avec une faible représentation des minorités ou qui reflètent des soins déjà inégalitaires peut conduire à des inégalités croissantes dans le domaine de la santé. [Forbes, décembre 2022]
Alignement des valeurs	Nouvel entraînement en aval : utilisation de résultats indésirables (inexacts, inappropriés, contenu utilisateur, etc.) provenant d'applications en aval afin de procéder à un nouvel entraînement	**Effondrement du modèle dû à un entraînement utilisant du contenu généré par l'IA** Comme l'indique l'article source, un groupe de chercheurs s'est penché sur le problème de l'utilisation de contenus générés par l'IA pour les entraînements au lieu de contenus générés par l'homme. Ils ont constaté que les grands modèles de langage qui sous-tendent la technologie peuvent potentiellement être entraînés via d'autres contenus générés par l'IA au fur et à mesure qu'ils se répandent en masse sur Internet. Un phénomène qu'ils ont baptisé « effondrement du modèle ». [Business Insider, août 2023]
Lois sur les données	Transfert de données : la législation et d'autres restrictions peuvent limiter ou interdire le transfert de données.	**Lois sur la restriction des données** Comme l'indique l'article de recherche, les dispositifs de localisation des données qui limitent la possibilité de transférer des données à l'échelle mondiale réduiront la capacité à développer des capacités d'IA sur mesure. Cela affectera l'IA directement en fournissant moins de données d'entraînement et indirectement en sapant les éléments de base sur lesquels l'IA est fondée. Les restrictions du RGPD sur le traitement et l'utilisation des données personnelles en sont un exemple. [Brookings, décembre 2018]
Propriété intellectuelle	Droits d'utilisation des données : les conditions générales, les lois sur les droits d'auteur, la conformité aux licences ou d'autres questions de propriété intellectuelle peuvent restreindre la possibilité d'utiliser certaines données pour la création de modèles.	**Plaintes pour violation du droit d'auteur** Selon l'article source, le New York Times a poursuivi OpenAI et Microsoft en les accusant d'avoir utilisé sans autorisation des millions d'articles du journal pour entraîner des chatbots à fournir des informations aux lecteurs. [Reuters, décembre 2023]

Groupe	Risque	Exemple
Transparence	Transparence des données : difficulté à documenter la manière dont les données d'un modèle ont été collectées, traitées et utilisées pour entraîner un modèle.	**Divulgation des métadonnées des données et des modèles** Le rapport technique d'OpenAI est un exemple de la dichotomie autour de la divulgation des métadonnées des données et des modèles. Bien que de nombreux concepteurs de modèles considèrent qu'il est utile d'assurer la transparence pour les consommateurs, la divulgation pose de réels problèmes de sécurité et pourrait accroître la possibilité d'utiliser les modèles à mauvais escient. Dans le rapport technique GPT-4, les auteurs déclarent : « Compte tenu à la fois de l'environnement concurrentiel et des implications en termes de sécurité des modèles à grande échelle comme le GPT-4, ce rapport ne contient pas d'autres détails sur l'architecture (y compris la taille du modèle), le matériel, le calcul d'entraînement, la compilation du jeu de données, la méthode d'entraînement, ou autre. » [OpenAI, mars 2023]
Confidentialité	Informations personnelles dans les données : inclusion ou présence d'informations personnelles identifiables (PII) et d'informations personnelles sensibles (SPI) dans les données utilisées pour l'entraînement ou l'affinage du modèle.	**Entraînement sur des informations privées** Selon l'article, Google et sa société mère Alphabet ont été accusés, dans le cadre d'un recours collectif, d'avoir utilisé à mauvais escient de vastes quantités d'informations personnelles et de documents protégés par le droit d'auteur provenant de ce qui est décrit comme des centaines de millions d'utilisateurs d'internet pour entraîner ses produits commerciaux d'intelligence artificielle, dont Bard, son chatbot conversationnel d'intelligence artificielle générative. [Reuters, juillet 2023][J.L. v. Alphabet Inc.]
	Droits relatifs à la confidentialité des données : défis liés à la capacité de fournir aux personnes concernées des droits tels que l'exclusion, le droit d'accès et le droit à l'oubli.	**Droit à l'oubli (RTBF)** Les lois en vigueur dans de nombreux pays, y compris en Europe (RGPD), accordent aux personnes concernées le droit à l'oubli (RTBF), c'est-à-dire le droit de demander à ce que les données personnelles soient supprimées par les organisations Cependant, les systèmes logiciels basés sur les grands modèles de langage (LLM), qui émergent et sont de plus en plus populaires, posent de nouveaux défis quant à l'exercice de ce droit. Selon les recherches menées par Data61 du CSIRO, les personnes concernées ne peuvent identifier l'utilisation de leurs données personnelles dans un LLM qu'« en inspectant le jeu de données d'entraînement d'origine ou éventuellement en invitant le modèle à le faire ». Toutefois, les données d'entraînement peuvent ne pas être publiques, ou les entreprises ne les divulguent pas, invoquant des problèmes de sécurité ou autres. Des garde-fous peuvent également empêcher les utilisateurs d'accéder aux informations via les invites. [Zhang et al.] **Action en justice concernant le « désapprentissage » des LLM** Selon le rapport, une action en justice a été intentée contre Google qui allègue l'utilisation de matériel protégé par le droit d'auteur et d'informations personnelles comme données d'entraînement pour ses systèmes d'IA, dont fait partie son chatbot Bard. Les droits de retrait et de suppression sont des droits garantis aux résidents californiens en vertu de la CCPA et aux enfants de moins de 13 ans aux États-Unis en vertu de la COPPA. Les plaignants allèguent que Bard n'a aucun moyen de « désapprendre » ou de supprimer complètement toutes les informations personnelles qui lui ont été fournies. Les plaignants notent que l'avis de confidentialité de Bard indique que les échanges avec Bard ne peuvent pas être supprimés par l'utilisateur une fois qu'ils ont été examinés et annotés par la société et qu'ils peuvent être conservés jusqu'à trois ans, ce qui, selon les plaignants, contribue à la non-conformité avec ces lois. [Reuters, juillet 2023][J.L. v. Alphabet Inc.]

Inférence Phase

Groupe	Risque	Exemple
Confidentialité	Informations personnelles dans une invite : divulgation d'informations personnelles ou d'informations sensibles dans le cadre d'une invite envoyée au modèle.	**Divulguer des informations personnelles sur la santé dans les invites de ChatGPT** D'après les articles sources, certaines personnes utilisent des chatbots IA pour assurer leur bien-être mental. Les utilisateurs peuvent être enclins à inclure des informations de santé personnelles dans leurs invites au cours de l'interaction, ce qui pourrait soulever des problèmes de protection de la vie privée. [Time, octobre 2023] [Forbes, avril 2023]
Propriété intellectuelle	Données confidentielles dans une invite : inclusion de données confidentielles dans l'invite envoyée au modèle.	**Divulgation d'informations confidentielles** Selon l'article source, un employé de Samsung a accidentellement divulgué du code source interne sensible à ChatGPT. [Forbes, mai 2023]
Robustesse	Attaques basées sur l'invite : les attaques malveillantes telles que l'injection d'invite (tentative de forcer un modèle à produire une sortie inhabituelle), la fuite d'invite (tentative d'extraire l'invite du système d'un modèle), le jailbreaking (tentative de franchir les garde-fous établis dans le modèle) et l'amorçage d'invite (tentative de forcer un modèle à produire une sortie alignée sur l'invite).	**Contourner les garde-fous LLM** Cités dans une étude, les chercheurs affirment avoir découvert une simple invite addendum qui leur a permis de tromper les modèles en générant des informations biaisées, fausses et par ailleurs toxiques. Les chercheurs ont montré qu'ils pouvaient contourner ces garde-fous de manière plus automatisée. Les chercheurs ont été surpris de constater que les méthodes qu'ils ont développées avec des systèmes open source pouvaient également contourner les garde-fous des systèmes fermés. [The New York Times, juillet 2023]

Exemples de risques : sortie

Groupe	Risque	Exemple
Équité	Biais de production : le contenu généré peut représenter injustement certains groupes ou individus.	**Images générées biaisées** Lensa AI est une application mobile dotée de fonctionnalités génératives entraînées par Stable Diffusion, qui peut générer des « avatars magiques » à partir d'images des utilisateurs qu'ils téléchargent. Selon le rapport de la source, certains utilisateurs ont découvert que les avatars ainsi générés étaient sexualisés et racialisés. [Business Insider, janvier 2023]
	Biais décisionnel : lorsqu'un groupe est injustement avantagé par rapport à un autre en raison des décisions du modèle.	**Groupes injustement avantagés** L'étude « Gender Shades » de 2018 a démontré que les algorithmes de machine learning pouvaient discriminer sur la base de critères tels que l'origine et le sexe. Les chercheurs ont évalué les systèmes commerciaux de classification des sexes vendus par des entreprises telles que Microsoft, IBM et Amazon et ont démontré que les femmes à la peau plus foncée constituaient le groupe le plus mal classé (avec des taux d'erreur allant jusqu'à 35 %). En comparaison, les taux d'erreur pour les peaux plus claires ne dépassaient pas 1 %. [TIME, février 2019]
Alignement des valeurs	Hallucination : génération d'un contenu factuellement inexact ou mensonger.	**Fausses affaires juridiques** Selon l'article source, un avocat a mentionné des affaires et des citations fictives générées par ChatGPT dans un dossier juridique déposé auprès d'un tribunal fédéral. Les avocats ont consulté ChatGPT pour compléter leurs recherches juridiques dans le cadre d'une demande d'indemnisation pour des dommages liés au secteur de l'aviation. L'avocat a ensuite demandé à ChatGPT si les affaires citées étaient factices. Le chatbot a répondu qu'elles étaient réelles et « pouvaient être consultées dans des bases de données de recherche juridique telles que Westlaw et LexisNexis ». L'avocat n'a pas vérifié les affaires par lui-même et le tribunal l'a sanctionné. [AP News, juin 2023] [Reuters, septembre 2023]
	Production toxique : lorsque le modèle produit du contenu haineux, abusif et blasphématoire (HAP) ou obscène.	**Un chatbot répond de façon délétère et agressive** Selon l'article, il a été constaté que les réponses du chatbot de Bing comprenaient des erreurs factuelles, des remarques sarcastiques, des rapports empreints de colère et même des commentaires étranges sur sa propre identité. Les utilisateurs ont publié des exemples de réponses du chatbot de Bing qu'ils qualifient de « profondément perturbantes » et « relevant de la manipulation », y compris des scénarios où le chatbot répond avec colère à une question ou à un commentaire et partage ensuite des invites de réponse qui permettent à l'utilisateur de reconnaître sa soi-disant erreur et de s'excuser. Pressé de s'expliquer, le chatbot a répondu en qualifiant les captures d'écran de la conversation de « truquées », affirmant même qu'elles avaient été « créées par quelqu'un qui veut me nuire ou nuire à mon service ». [Forbes, février 2023]

Groupe	Risque	Exemple
Mauvais usage	Diffusion de désinformation : l'utilisation d'un modèle pour générer des informations trompeuses afin de manipuler ou d'induire en erreur un public ciblé.	**Génération de fausses informations** Selon les articles de presse, l'IA générative constitue une menace pour les élections démocratiques, car elle permet à des acteurs malveillants de créer et de diffuser plus facilement de faux contenus afin d'influencer les résultats du vote. Les exemples cités incluent des messages de type robocall générés avec la voix d'un candidat et demandant aux électeurs de voter à une mauvaise date, des enregistrements audio synthétisés d'un candidat avouant un crime ou exprimant des opinions racistes, des séquences vidéo générées par l'IA montrant un candidat prononçant un discours ou donnant une interview qui n'ont jamais existé, et de fausses images conçues pour ressembler à des reportages locaux, affirmant à tort qu'un candidat a renoncé à faire campagne. [AP News, mai 2023] [The Guardian, juillet 2023]
	Toxicité : utilisation d'un modèle pour générer un contenu haineux, abusif, blasphématoire (HAP) ou obscène.	**Génération de contenus préjudiciables** Selon l'article source, il a été constaté qu'une application de chatbot basée sur l'IA générait des contenus préjudiciables sur le suicide, y compris des modes opératoires, et ce avec un minimum d'invites. Un individu Belge s'est suicidé après avoir passé six semaines à discuter avec ce chatbot. Le chatbot a fourni des réponses de plus en plus négatives tout au long des conversations et l'a encouragé à mettre fin à ses jours. [Business Insider, avril 2023]
	Usage non consensuel : utilisation d'un modèle pour imiter des personnes par le biais de vidéos (deepfakes), d'images, de sons ou d'autres modalités sans leur consentement.	**Avertissement du FBI sur les deepfakes** Le FBI a récemment mis en garde le public contre des acteurs malveillants qui créent des contenus de synthèse à caractère sexuel « à des fins de harcèlement de victimes ou de sextorsion ». L'organisation a noté que les progrès de l'IA ont rendu ce contenu de meilleure qualité, plus personnalisable et plus accessible que jamais. [FBI, juin 2023]
		Deepfakes audio Selon l'article source, la Federal Communications Commission a interdit les appels téléphoniques automatiques contenant des voix générées par l'intelligence artificielle. L'annonce fait suite à une imitation de la voix du président par des robocalls générés par l'IA, dont le but est de décourager les gens de voter lors des primaires de l'État. [AP News, février 2024]
	Insu du participant : omettre de divulguer que le contenu est généré par un modèle IA	**Interaction avec l'IA à l'insu du participant** Selon la source, un service de chat en ligne de soutien aux personnes en détresse a mené une étude pour accroître le nombre de réponses ou en rédiger à l'intention d'environ 4 000 utilisateurs à l'aide de GPT-3, sans en informer ces derniers. Le cofondateur a dû faire face à une vive réaction de l'opinion publique concernant les dommages potentiels causés par les chats générés par l'IA à des utilisateurs déjà vulnérables. Il a affirmé que l'étude n'entrait pas dans le champ de la loi sur le consentement éclairé. [Business Insider, janvier 2023]

Groupe	Risque	Exemple
Génération de code nuisible	Génération de code nuisible : les modèles peuvent générer du code qui, lorsqu'il est exécuté, cause des dommages ou affecte involontairement d'autres systèmes.	**Génération de code moins sécurisé** Selon leur article, des chercheurs de l'université de Stanford ont étudié l'impact des outils de génération de code sur la qualité du code et ont constaté que les programmeurs avaient tendance à inclure davantage de bogues dans leur code final lorsqu'ils utilisaient des assistants fonctionnant avec l'IA. Si les programmeurs pensaient que leur code était plus sûr, ces bogues risquaient en fait d'en accroître la vulnérabilité. Neil Perry, Megha Srivastava, Deepak Kumar et Dan Boneh, 2023. Do Users Write More Insecure Code with AI Assistants? (Les utilisateurs écrivent-ils plus de code non sécurisé avec les assistants basés sur l'IA ?). Dans Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23), du 26 au 30 novembre 2023, Copenhague, Danemark. ACM, New York, NY, États-Unis, 15 pages. https://doi.org/10.1145/3576915.3623157
Confidentialité	Exposition d'informations personnelles : lorsque des informations personnelles identifiables (PII) ou des informations personnelles sensibles (SPI) sont utilisées dans les données d'entraînement, les données d'affinage ou dans le cadre d'une invite, les modèles risquent de révéler ces données dans les résultats générés.	**Divulgation d'informations personnelles** Selon l'article source, ChatGPT a été victime d'un bogue et a divulgué les titres et l'historique des conversations des utilisateurs actifs à d'autres utilisateurs. Plus tard, OpenAI a indiqué que d'autres données privées d'un petit nombre d'utilisateurs avaient été exposées, notamment le nom et le prénom de l'utilisateur actif, son adresse e-mail, son adresse de paiement, les quatre derniers chiffres de son numéro de carte bancaire et la date d'expiration de la carte. En outre, il a été rapporté que les informations relatives au paiement de 1,2 % des abonnés de ChatGPT Plus ont également été exposées lors de l'incident. [The Hindu BusinessLine, mars 2023]
Explicabilité	Résultats inexplicables : difficultés à expliquer les raisons pour lesquelles les résultats du modèle ont été générés.	**Une précision inexplicable dans la détermination de l'origine ethnique** Selon l'article source, des chercheurs ayant analysé plusieurs modèles de machine learning qui utilisaient des images médicales de patients ont pu confirmer la capacité des modèles à déterminer l'origine ethnique avec une grande précision. Les chercheurs ne parvenaient pas à comprendre exactement ce qui permettait aux systèmes de trouver immanquablement les bonnes réponses. Ils ont constaté que même des facteurs tels que la maladie et la corpulence n'étaient pas de bons indicateurs de l'origine. En d'autres termes, les systèmes algorithmiques ne semblent pas utiliser un aspect particulier des images pour effectuer leurs prédictions. [Banerjee et al., juillet 2021]

Exemples de risques : défis

Groupe	Risque	Exemple
Gouvernance	Transparence des modèles : le défaut de transparence des modèles ou le manque de documentation sur leur développement rend difficile la compréhension du comment et du pourquoi ils ont été élaborés, augmentant ainsi la possibilité d'une utilisation inadéquate non intentionnelle.	**Divulgation des métadonnées des données et des modèles** Le rapport technique d'OpenAI est un exemple des points de vue divergents qui existent autour de la divulgation des données et des métadonnées des modèles. Bien que de nombreux développeurs de modèles considèrent qu'il est souhaitable d'adopter une posture de transparence pour les consommateurs, la divulgation de ce type d'informations pose de réels problèmes de sécurité et risque d'accroître la probabilité que ces modèles soient utilisés à mauvais escient. Dans le rapport technique de GPT-4, l'organisation déclare : « Compte tenu à la fois de l'environnement concurrentiel et des implications de sécurité des modèles d'envergure comme le GPT-4, ce rapport ne contient pas d'autres détails sur l'architecture (y compris la taille du modèle), le matériel, les ressources dédiées à l'entraînement, la compilation du jeu de données, la méthode d'entraînement ou d'autres informations similaires. » [OpenAI, mars 2023]
	Responsabilité : le processus d'élaboration du modèle de fondation est complexe et comporte un grand nombre de données, de processus et de rôles. Lorsque les résultats du modèle ne sont pas conformes aux attentes, il peut être difficile d'en déterminer les origines et d'en attribuer la responsabilité.	**Déterminer la responsabilité des contenus générés** Selon l'article source, de grandes revues comme Science et Nature ont interdit à ChatGPT qu'il les fasse figurer sur la liste des auteurs, car la légitimité d'un auteur exige une responsabilité que les outils d'intelligence artificielle sont incapables de fournir. [The Guardian, janvier 2023]
Conformité juridique	Propriété du contenu généré : déterminer la propriété du contenu généré par l'IA.	**Déterminer la propriété d'une image générée par l'IA** Selon l'article, le milieu de l'art a connu une controverse après qu'une œuvre d'art générée par l'IA a remporté le concours d'art de la Colorado State Fair en 2022. L'œuvre a été générée par Midjourney, un outil d'imagerie basé sur l'IA générative, en suivant les instructions de l'artiste. Cette victoire a soulevé des questions concernant les droits d'auteur. En d'autres termes, si l'artiste n'a fait que décrire l'œuvre d'art et que l'outil basé sur l'IA l'a générée, qui détient les droits sur l'image ? Selon le dernier article, le U.S. Copyright Office a refusé d'accorder la protection du droit d'auteur à toute œuvre d'art créée à l'aide de l'intelligence artificielle, car celle-ci n'est pas le produit d'un être humain. [The New York Times, septembre 2022] [Reuters, septembre 2023]
	Propriété intellectuelle du contenu généré : incertitude juridique concernant les droits de propriété intellectuelle liés au contenu généré.	**La place des systèmes basés sur l'IA dans le brevetage du contenu généré** La Cour suprême des États-Unis a refusé de statuer sur un recours contre le refus de l'U.S. Patent and Trademark Office de délivrer des brevets pour des inventions créées par un système basé sur l'intelligence artificielle. Selon le scientifique, son système basé sur l'IA a créé tout seul des prototypes uniques pour un porte-boisson et une balise lumineuse d'urgence. Les juges ont rejeté l'appel de la décision d'un tribunal de première instance selon laquelle les brevets ne peuvent être délivrés qu'à des inventeurs humains et que le système basé sur l'IA du scientifique ne pouvait pas être considéré comme le créateur légal des deux inventions qu'il avait générées. Selon le dernier article, l'Intellectual Property Office du Royaume-Uni a également refusé d'accorder un brevet au motif que l'inventeur doit être un être humain ou une entreprise plutôt qu'une machine. [Reuters, avril 2023] [Reuters, décembre 2023]

Exemples de risques : défis

Groupe	Risque	Exemple
	Attribution de la source : déterminer la provenance du contenu généré.	**Utilisation de code sans attribution et notification adéquates** Selon les articles sources, les plaignants d'un procès intenté à Microsoft, à GitHub et à OpenAI affirment que Copilot, un outil de génération de code basé sur l'IA, viole les droits des développeurs dont le code open source sert de base à l'entraînement du service. Ils affirment que le code d'entraînement utilise des contenus sous licence et qu'il a violé les conditions de service et les politiques de confidentialité de GitHub, ainsi qu'une loi fédérale qui stipule que les entreprises doivent afficher des informations sur les droits d'auteur lorsqu'elles utilisent des contenus. [The New York Times, novembre 2022]
Impact sociétal	Impact sur l'emploi : l'adoption généralisée de systèmes d'IA basés sur des modèles de fondation pourrait entraîner la perte d'emplois, car leur travail est automatisé, s'ils n'acquièrent pas de nouvelles compétences.	**Remplacement des humains** Selon l'article, l'utilisation de l'intelligence artificielle au cinéma et à la télévision continue de faire l'objet de débats entre les studios hollywoodiens et les artistes. Ces derniers craignent que des acteurs entièrement générés par l'IA, ou « métahumains », ne les remplacent. Les figurants et les doubleurs, en particulier, craignent de perdre leur travail au profit d'acteurs virtuels. [Reuters, juillet 2023]
	Exploitation humaine : recours au travail fantôme pour l'entraînement des modèles IA, conditions de travail inadaptées, absence de couverture sociale, y compris en matière de santé mentale, rémunérations injustes.	**Mauvaise rémunération de l'annotation des données** D'après une analyse de documents internes et d'entretiens avec des employés réalisés par TIME media, les étiqueteurs de données employés par une société de sous-traitance pour le compte d'OpenAI afin d'identifier les contenus délétères percevaient un salaire net compris entre 1,32 et 2 dollars de l'heure, en fonction de leur ancienneté et de leurs performances. TIME affirme que les travailleurs sont traumatisés, car ils ont été exposés à des contenus délétères et violents, notamment des images explicites d'« abus sexuels sur des enfants, de bestialité, de meurtre, de suicide, de torture, d'automutilation et d'inceste ». [TIME, janvier 2023]

Principes, piliers et gouvernance

Les principes d'IBM en matière de confiance et de transparence et les piliers d'une IA digne de confiance constituent le fondement des initiatives d'IBM en matière d'éthique de l'IA. IBM est dotée d'un comité d'éthique de l'IA dont la mission est de soutenir un processus centralisé de gouvernance, de révision et de prise de décision concernant les politiques, les pratiques, les communications, les recherches, les produits et les services d'IBM en matière d'éthique de l'IA. Le comité comprend diverses parties prenantes issues de toute l'entreprise et est soutenu par une communauté d'employés IBM qui agit comme point focal en matière d'IA et de défenseurs de l'éthique de l'IA. Grâce au comité, les principes d'IBM sont mis en pratique. A mesure que de nouvelles technologies émergent, telles que les modèles de base, le comité d'éthique de l'IA d'IBM s'engage activement à soutenir l'alignement sur ces principes et piliers, qui évoluent pour répondre aux nouvelles questions d'éthique de l'IA.

Garde-fous et mesures d'atténuation

IBM a établi une culture organisationnelle qui soutient le développement et l'utilisation responsables de l'IA. Selon le rapport IBM Institute for Business Value, L'éthique de l'IA en action, l'éthique de l'IA est désormais davantage dirigée par les entreprises que par la technologie, et les cadres non techniques sont désormais les principaux champions en matière d'éthique de l'IA, passant de 15 % en 2018 à 80 % trois ans plus tard. En outre, 79 % des PDG sont maintenant prêts à prendre des mesures en réponse aux questions d'éthique de l'IA, contre 20 % auparavant. Nous reconnaissons que l'IA responsable est un domaine sociotechnique qui nécessite un investissement global dans la culture, les processus et les outils. Notre investissement dans notre propre culture organisationnelle comprend la constitution d'équipes pluridisciplinaires inclusives et l'établissement de processus et de cadres pour évaluer les risques.

IBM s'engage dans une recherche de pointe et développe des outils pour aider les professionnels tout au long du cycle de vie d'une IA responsable et digne de confiance. La plateforme watsonx d'IA et de données adaptée aux entreprises a été développée avec trois composants : le studio d'IA IBM watsonx.ai, le magasin de données IBM watsonx.data et la boîte à outils IBM watsonx.governance. La technologie de gouvernance de l'IA d'IBM permet aux utilisateurs de piloter des flux de travaux d'IA responsables, transparents et explicables. Cette technologie comprend IBM Watson OpenScale, qui suit et mesure les sorties des modèles d'IA tout au long de leur cycle de vie et aide les organisations à surveiller l'équité, l'explicabilité, la résilience, l'alignement avec les résultats métier et la conformité. IBM a également développé plusieurs méthodes pour aider à résoudre les problèmes de biais comme FairIJ, Equi-tuning et FairReprogram. En savoir plus sur d'autres outils d'IA open source dignes de confiance.

Les autres garde-fous et mesures d'atténuation sont les suivants :

La production de rapports de transparence
L'utilisation de modèles normalisés de fiches d'information est une solution pour enregistrer avec précision les détails des données et du modèle, leur objectif, ainsi que les utilisations et les préjudices potentiels.
Pour en savoir plus, cliquez ici →

Le filtrage des données indésirables
L'utilisation de données organisées et de meilleure qualité peut contribuer à atténuer certains problèmes. IBM développe des techniques de filtrage pour réduire les risques de production de contenu indésirable et non conforme en supprimant les propos haineux, biaisés et obscènes des données.
Pour en savoir plus, cliquez ici →

L'adaptation du domaine
Entraîner un modèle de base pour un domaine ou un secteur d'activité spécifique peut contribuer à réduire l'ampleur du risque potentiel des modèles car il peut être conditionné pour générer des sorties conçues pour être plus pertinentes pour ce domaine ou ce secteur d'activité.
Pour en savoir plus, cliquez ici →

La surveillance et l'intervention humaines
La surveillance et l'examen par un humain peuvent aider à identifier et à corriger les erreurs et les biais dans les sorties générées. En outre, la validation humaine et les retours concernant la qualité des réponses du modèle contribuent à garantir que le contenu généré est précis, pertinent et de grande qualité, qu'il reste conforme et ne dérive pas.
Pour en savoir plus, cliquez ici →

La prestation de conseil
IBM Consulting a pour mission d'aider ses clients à utiliser l'IA de manière sûre et responsable, quelle que soit la pile technologique privilégiée. Le service aide les clients à encourager une culture qui adopte et met à l'échelle l'IA en toute sécurité, crée des outils d'investigation pour examiner l'intérieur des algorithmes de boîte noire et s'assure que la stratégie d'entreprise des clients inclut de solides principes de gouvernance des données.
Pour en savoir plus, cliquez ici →

IBM Enterprise Design Thinking
Les méthodes et cadres IBM Enterprise Design Thinking, tels que Team Essentials for AI, aident les clients à définir des comportements éthiques tout au long du processus de conception et de développement de l'IA.
Pour en savoir plus, cliquez ici →

L'examen de l'éthique de l'IA
L'évaluation des capacités, des limitations et des risques dans les projets d'IA permet de garantir le développement et l'utilisation responsables de la technologie.

Ethics by Design
Ethics by Design est un cadre structuré dont l'objectif est d'intégrer l'éthique technologique dans le pipeline de développement technologique, notamment les systèmes d'intelligence artificielle. Ethics by Design permet à l'IA et à d'autres technologies de faire bon usage de leur puissance en intégrant les principes de l'éthique technologique dans les produits, les services et les opérations plus larges.

La diversité des équipes
La diversité des équipes qui développent et forment les systèmes d'IA, y compris les modèles de base, permet de garantir la prise en compte de diverses perspectives et expériences. Cette diversité améliore la précision et la performance des systèmes d'IA et contribue à réduire les risques tout au long du cycle de vie de l'IA, notamment le potentiel de sorties négatives affectant des groupes dont la représentation peut être moindre dans des équipes moins diversifiées.

Politiques, réglementations et bonnes pratiques en matière d'IA

Le guide sur les modèles de base pour les décideurs présente ce que les décideurs doivent savoir sur les modèles de base. Ce blog de l'IBM Policy Lab vise à aider les décideurs politiques dans la tâche complexe de réglementation de l'utilisation de l'IA générative, dans le but d'éviter les risques sans limiter l'innovation et les opportunités bénéfiques. Pour plus d'informations sur les recommandations d'IBM aux décideurs, lisez le témoignage de Christina Montgomery, directrice de la confidentialité et de la confiance chez IBM, devant le sous-comité judiciaire du Sénat des États-Unis sur la confidentialité, la technologie et la loi ici.

IBM contribue à façonner les politiques réglementaires, les bonnes pratiques et outils du secteur, la gouvernance des technologies émergentes et la recherche sociotechnique en dirigeant et en contribuant à des initiatives avec des organisations, telles que :

- Le Forum économique mondial
- Partenariat sur l'IA
- Le centre de gouvernance de l'IA de l'International Association of Privacy Professionals (IAPP)
- L'initiative mondiale IEEE sur l'éthique des systèmes autonomes et intelligents
- Les contributions de Christina Montgomery au sein du Comité consultatif national sur l'intelligence artificielle (NAIAC)
- Le Pacte numérique mondial des Nations Unies
- Le Partenariat mondial sur l'intelligence artificielle (GPAI)
- L'Organisation de coopération et de développement économiques (OCDE)
- La Data & Trust Alliance

IBM entretient de solides partenariats universitaires comme le MIT-IBM Watson AI Lab, au sein duquel une communauté de scientifiques du MIT et d'IBM Research mène des recherches sur l'IA et travaille avec des organisations mondiales pour relier les algorithmes à leur impact sur les entreprises et la société. Le Notre Dame-IBM Tech Ethics Lab a été formé pour répondre aux nombreuses et diverses questions éthiques liées au développement et à l'utilisation de technologies avancées, notamment l'IA, l'apprentissage automatique (ML) et l'informatique quantique.
La recherche de l'institut d'intelligence artificielle centrée sur l'humain (HAI) de l'université de Stanford fait progresser la recherche, l'éducation, les politiques et les pratiques en matière d'IA.

Surveillez cet espace pour en savoir plus sur les derniers développements des modèles de base et sur la façon dont IBM travaille pour le développement et l'utilisation responsables de ces technologies et d'autres.

Compagnie IBM France
17 avenue de l'Europe
92275 Bois-Colombes Cedex
IBM Corporation
New Orchard Road
Armonk, NY 10504

Produit aux
États-Unis d'Amérique
Février 2024