NIST AI 100-1

Intelligence artificielle Gestion des risques (AI RMF 1.0)

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST AI 100-1

Intelligence artificielle Gestion des risques (AI RMF 1.0)

Cette publication est disponible gratuitement à l'adresse suivante :
https://doi.org/10.6028/NIST.AI.100-1

janvier 2023

Département du commerce des États-Unis
Gina M. Raimondo, Secrétaire

Institut national des normes et de la technologie
Laurie E. Locascio, directeur du NIST et sous-secrétaire au commerce pour les normes et la technologie

Certaines entités commerciales, certains équipements ou matériaux peuvent être identifiés dans ce document afin de décrire correctement une procédure ou un concept expérimental. Cette identification n'a pas pour but de recommander ou d'approuver le National Institute of Standards and Technology, ni d'impliquer que les entités, les matériaux ou l'équipement sont nécessairement les meilleurs disponibles pour l'objectif visé.

Cette publication est disponible gratuitement à l'adresse suivante : https://doi.org/10.6028/NIST.AI.100-1

Calendrier des mises à jour et versions

Le cadre de gestion des risques liés à l'intelligence artificielle est un document évolutif.

Le NIST réexaminera régulièrement le contenu et l'utilité du cadre afin de déterminer si une mise à jour est ; une révision avec une contribution formelle de la communauté de l'IA devrait avoir lieu au plus tard en 2028. Le cadre utilisera un système de versions à deux numéros pour suivre et identifier les changements majeurs et mineurs. Le premier numéro représentera la génération du cadre de référence de l'IA et des documents qui l'accompagnent (par exemple, 1.0) et ne changera qu'en cas de révisions majeures. Les révisions mineures seront identifiées en utilisant ".n" après le numéro de génération (par exemple, 1.1). Toutes les modifications seront suivies à l'aide d'un tableau de contrôle des versions qui identifie l'historique, y compris le numéro de version, la date de modification et la description de la modification. Le NIST prévoit de mettre fréquemment à jour le manuel AI RMF. Les commentaires sur le AI RMF Playbook peuvent être envoyés par courriel àAIframework@nist.gov à tout et seront examinés et intégrés sur une base semestrielle.

Table des matières

Liste des tableaux

Liste des figures

Résumé

Les technologies de l'intelligence artificielle (IA) ont un potentiel considérable pour transformer la société et la vie des gens - du commerce et de la santé aux transports et à la cybersécurité, en passant par l'environnement et notre planète. Les technologies de l'IA peuvent favoriser une croissance économique inclusive et soutenir les avancées scientifiques qui améliorent les conditions de vie dans notre monde. Toutefois, les technologies de l'IA présentent également des risques susceptibles d'avoir un impact négatif sur les individus, les groupes, les organisations, les communautés, la société, l'environnement et la planète. À l'instar des risques liés à d'autres types de technologies, les risques liés à l'IA peuvent se manifester de différentes manières et peuvent être caractérisés comme étant à long ou à court terme, à forte ou à faible probabilité, systémiques ou localisés, et à fort ou à faible impact.

Le cadre de référence de l'IA fait référence à un *système d'IA* en tant que système technique ou mécanique qui peut, pour un ensemble donné d'objectifs, générer des résultats tels que des prédictions, des recommandations ou des décisions influençant des environnements réels ou virtuels. Les systèmes d'IA sont conçus pour fonctionner avec différents niveaux d'autonomie (adapté de : Recommandation de l'OCDE sur l'IA:2019 ; ISO/IEC 22989:2022).

S'il existe une multitude de normes et de bonnes pratiques pour aider les organisations à atténuer les risques liés aux logiciels traditionnels ou aux systèmes basés sur l'information, les risques posés par les systèmes d'IA sont bien des égards uniques (voir l'annexe B). Les systèmes d'IA, par exemple, peuvent être formés à partir de données qui peuvent changer au fil du temps, parfois de manière significative et inattendue, ce qui affecte le fonctionnement et la fiabilité du système d'une manière difficile à comprendre. Les systèmes d'IA et les contextes dans lesquels ils sont déployés sont souvent complexes, ce qui rend difficile la détection des défaillances et la réaction à celles-ci lorsqu'elles se produisent. Les systèmes d'IA sont par nature sociotechniques, ce qui signifie qu'ils sont influencés par la dynamique sociétale et le comportement humain. Les risques - et les avantages - de l'IA peuvent résulter de l'interaction entre des aspects techniques et des facteurs sociétaux liés à la manière dont un système est utilisé, à ses interactions avec d'autres systèmes d'IA, aux personnes qui l'exploitent et au contexte social dans lequel il est déployé.

Ces risques font de l'IA une technologie particulièrement difficile à déployer et à utiliser, tant pour les organisations que pour la société. Sans contrôles appropriés, les systèmes d'IA peuvent amplifier, perpétuer ou exacerber des résultats inéquitables ou indésirables pour les individus et les communautés. Avec des contrôles appropriés, les systèmes d'IA peuvent atténuer et gérer les résultats inéquitables.

La gestion des risques liés à l'IA est un élément clé du développement et de l'utilisation responsables des systèmes d'IA. Les pratiques d'IA responsable peuvent contribuer à aligner les décisions relatives à la conception, au développement et à l'utilisation des systèmes d'IA sur les objectifs et les valeurs visés. Les concepts fondamentaux de l'IA responsable mettent l'accent sur l'humain, la responsabilité sociale et la durabilité. La gestion des risques liés à l'IA peut favoriser des utilisations et des pratiques responsables en incitant les organisations et leurs équipes internes qui conçoivent, développent et déploient l'IA à réfléchir de manière plus critique au contexte et aux impacts négatifs et positifs potentiels ou inattendus. La compréhension et la gestion des risques liés aux systèmes d'IA contribueront à renforcer la fiabilité et, partant, à cultiver la confiance du public.

La responsabilité sociétale peut se référer à la responsabilité de l'organisation "pour les impacts de ses décisions et activités sur la société et l'environnement par un comportement transparent et éthique" (ISO 26000:2010). *La durabilité* fait référence à "l'état du système mondial, y compris les aspects environnementaux, sociaux et économiques, dans lequel les besoins du présent sont satisfaits sans compromettre la capacité des générations futures à satisfaire leurs propres besoins" (ISO/IEC TR 24368:2022). L'IA responsable est censée déboucher sur une technologie qui est également équitable et responsable. On attend des pratiques organisationnelles qu'elles soient conformes à la "*responsabilité professionnelle*", définie par l'ISO comme une approche qui "vise à garantir que les professionnels qui conçoivent, développent ou déploient des systèmes et applications d'IA ou des produits ou systèmes basés sur l'IA, reconnaissent leur position unique pour exercer une influence sur les personnes, la société et l'avenir de l'IA" (ISO/IEC TR 24368:2022).

Conformément à la National Artificial Intelligence Initiative Act of 2020 (P.L. 116-283), l'objectif de l'AI RMF est d'offrir une ressource aux organisations qui conçoivent, développent, déploient ou utilisent des systèmes d'IA afin de les aider à gérer les nombreux risques liés à l'IA et de promouvoir un développement et une utilisation confiants et responsables des systèmes d'IA. Le cadre se veut ***volontaire***, respectueux des droits, non spécifique à un secteur et indépendant des cas d'utilisation, offrant ainsi aux organisations de toute taille, de tout secteur et de toute la société la possibilité de mettre en œuvre les approches qu'il contient.

Le cadre est conçu pour doter les organisations et les individus - appelés ici *acteurs de l'IA* - d'approches permettant d'accroître la fiabilité des systèmes d'IA et de favoriser la conception, le développement, le déploiement et l'utilisation responsables des systèmes d'IA au fil du temps. Les acteurs de l'IA sont définis par l'Organisation de coopération et de développement économiques (OCDE) comme "ceux qui jouent un rôle actif dans le cycle de vie des systèmes d'IA, y compris les organisations et les individus qui déploient ou exploitent l'IA" [OCDE (2019) L'intelligence artificielle dans la société - iLibrary de l'OCDE] (voir l'annexe A).

Le RMF est conçu pour être pratique, pour s'adapter au paysage de l'IA au fur et à mesure que les technologies de l'IA continuent à se développer, et pour être mis en œuvre par des organisations à des degrés et avec des capacités variables afin que la société puisse bénéficier de l'IA tout en étant protégée de ses inconvénients potentiels.

Le cadre et les ressources qui l'accompagnent seront mis à jour, étendus et améliorés en fonction de l'évolution de la technologie, du paysage des normes dans le monde et de l'expérience et du retour d'information de la communauté de l'IALe NIST continuera d'aligner le cadre de référence de l'IA et les orientations connexes sur les normes, lignes directrices et pratiques internationales applicables. Au fur et à mesure de l'utilisation du RMF, des enseignements supplémentaires seront tirés afin d'éclairer les futures mises à jour et les ressources additionnelles.

Le cadre est divisé en deux parties. La première partie explique comment les organisations peuvent encadrer les risques liés à l'IA et décrit le public visé. Ensuite, les risques et la fiabilité de l'IA sont analysés, en soulignant les caractéristiques des systèmes d'IA dignes de confiance, à savoir

valides et fiables, sûrs, sécurisés et résilients, responsables et transparents, explicables et interprétables, renforçant la protection de la vie privée, et équitables avec la gestion de leurs préjugés néfastes.

La partie 2 constitue le "cœur" du cadre. Elle décrit quatre fonctions spécifiques destinées à aider les organisations à gérer les risques liés aux systèmes d'IA dans la pratique. Ces fonctions - **GOUVERNER**, **CARTOGRAPHIER**, **MESURER** et **GÉRER** - sont subdivisées en catégories et sous-catégories. Alors que la **fonction GOUVERNER** s'applique à toutes les étapes des processus et procédures de gestion des risques liés à l'IA, les fonctions **MAP**, **MESURER** et **GÉRER** peuvent être appliquées dans des contextes spécifiques aux systèmes d'IA et à des étapes spécifiques du cycle de vie de l'IA.

Des ressources supplémentaires relatives au cadre sont incluses dans le AI RMF Playbook, qui est disponible sur le site web du NIST AI RMF :
https://www.nist.gov/itl/ai-risk-management-framework.

L'élaboration du cadre de référence pour l'IA par le NIST, en collaboration avec les secteurs privé et public, s'inscrit dans le cadre de ses efforts plus larges en matière d'IA, conformément à la loi de 2020 sur l'initiative ennationale matière d'IA, aux recommandations la commission de sécurité nationale sur l'intelligence artificielle de et au plan d'engagement fédéral pour l'élaboration de normes techniques et d'outils connexes. L'engagement de la communauté de l'IA au cours de l'élaboration de ce cadre - par le biais de réponses à une demande officielle d'informations, de trois ateliers très fréquentés, de commentaires publics sur un document de réflexion et deux versions préliminaires du cadre, de discussions lors de multiples forums publics et de nombreuses réunions en petits groupes - a contribué à l'élaboration du RMF 1.0 sur l'IA ainsi qu'à la recherche, au développement et à l'évaluation de l'IA menés par le NIST et par d'autres organismes. recherches prioritaires et les orientations supplémentaires qui amélioreront ce cadre seront consignées dans une feuille de route du cadre de gestion des risques liés à l'IA, à laquelle le NIST et l'ensemble de la communauté pourront contribuer.

Partie 1 : Informations de base

1. Encadrer le risque

La gestion des risques liés à l'IA permet de minimiser les effets négatifs potentiels des systèmes d'IA, tels que les menaces pour les libertés et les droits civils, tout en offrant la possibilité de maximiser les effets positifs. Le fait d'aborder, de documenter et de gérer efficacement les risques liés à l'IA et les effets négatifs potentiels peut conduire à des systèmes d'IA plus dignes de confiance.

1.1 Comprendre et traiter les risques, les impacts et les préjudices

Dans le contexte du CMR de l'IA, le *risque* désigne la mesure composite de la probabilité qu'un événement se produise et de l'ampleur ou du degré des conséquences de l'événement correspondant. Les impacts, ou conséquences, des systèmes d'IA peuvent être positifs, négatifs ou les deux à la fois et peuvent se traduire par des opportunités ou des menaces (Adapté de : ISO 31000:2018). Lorsqu'on considère l'impact négatif d'un événement potentiel, le risque est fonction 1) de l'impact négatif, ou de l'ampleur du préjudice, qui surviendrait si la circonstance ou l'événement se produisait et 2) de la probabilité d'occurrence (Adapté de : OMB Circular A-130:2016). L'impact négatif ou le préjudice peut être ressenti par les individus, les groupes, les communautés, les organisations, la société, l'environnement et la planète.

> "Le management du risque se réfère aux activités coordonnées pour diriger et contrôler une organisation en ce qui concerne le risque" (Source : ISO 31000:2018).

Alors que les processus de gestion des risques portent généralement sur les négatifs, le présent cadre propose des approches visant à minimiser les impacts négatifs anticipés des systèmes d'IA *et à* identifier les possibilités de maximiser les impacts positifs. Une gestion efficace des risques de dommages potentiels pourrait conduire à des systèmes d'IA plus fiables et libérer des avantages potentiels pour les personnes (individus, communautés et société), les organisations et les systèmes/écosystèmes. La gestion des risques peut permettre aux développeurs et aux utilisateurs de l'IA de comprendre les impacts et de tenir compte des limites et des incertitudes inhérentes à leurs modèles et à leurs systèmes, ce qui, à son tour, peut améliorer la performance et la fiabilité globales du système et la probabilité que les technologies de l'IA soient utilisées de manière bénéfique.

Le cadre réglementaire de l'IA est conçu pour faire face aux nouveaux risques au fur et à mesure qu'ils apparaissent. Cette flexibilité est particulièrement importante lorsque les impacts ne sont pas facilement prévisibles et que les applications évoluent. Si certains risques et avantages de l'IA sont bien connus, il peut être difficile d'évaluer les effets négatifs et l'ampleur des dommages. La figure 1 donne des exemples de préjudices potentiels liés aux systèmes d'IA.

Les efforts de gestion des risques liés à l'IA doivent tenir compte du fait que les humains peuvent supposer que les systèmes d'IA fonctionnent - et fonctionnent bien - dans *tous les* contextes. Par exemple, qu'ils soient corrects ou non, les systèmes d'IA sont souvent perçus comme étant plus objectifs que les humains ou comme offrant de plus grandes capacités que les logiciels généraux.

Fig. 1. Exemples de dommages potentiels liés aux systèmes d'IA. Des systèmes d'IA dignes de confiance et leur utilisation responsable peuvent atténuer les risques négatifs et contribuer aux avantages pour les personnes, les organisations et les écosystèmes.

1.2 Défis pour la gestion des risques liés à l'IA

Plusieurs défis sont décrits ci-dessous. Ils doivent être pris en compte lors de la gestion des risques dans le cadre de la recherche de la fiabilité de l'IA.

1.2.1 Mesure du risque

Les risques ou les défaillances de l'IA qui ne sont pas bien définis ou compris de manière adéquate sont difficiles à mesurer quantitativement ou qualitativement. L'incapacité à mesurer correctement les risques liés à l'IA ne signifie pas qu'un système d'IA présente nécessairement un risque élevé ou faible. Parmi les défis liés à la mesure des risques, on peut citer

Risques liés aux logiciels, au matériel et aux données de tiers : Les données ou systèmes de tiers peuvent accélérer la recherche et le développement et faciliter la transition technologique. Ils peuvent également compliquer la mesure des risques. Les risques peuvent provenir à la fois des données, des logiciels ou du matériel de tiers et de la manière dont ils sont utilisés. Les mesures ou méthodologies de risque utilisées par l'organisation qui développe le système d'IA peuvent ne pas correspondre aux mesures ou méthodologies de risque utilisées par l'organisation *qui déploie ou exploite* le système. En outre, l'organisation qui développe le système d'IA peut ne pas être transparente quant aux mesures de risque ou aux méthodologies qu'elle a utilisées. La mesure et la gestion des risques peuvent être compliquées par la manière dont les clients utilisent ou intègrent des données ou des systèmes de tiers dans les produits ou services d'IA, en particulier en l'absence de structures de gouvernance internes et de garanties techniques suffisantes. Quoi qu'il en soit, toutes les parties et tous les acteurs de l'IA devraient gérer les risques liés aux systèmes d'IA qu'ils développent, déploient ou utilisent en tant que composants autonomes ou intégrés.

Suivi des risques émergents : Les efforts de gestion des risques des organisations seront renforcés par l'identification et le suivi des risques émergents et par l'examen des techniques permettant de les mesurer.

Les approches d'évaluation de l'impact des systèmes d'IA peuvent aider les acteurs de l'IA à comprendre les impacts ou les préjudices potentiels dans des contextes spécifiques.

Disponibilité de mesures fiables : L'absence actuelle de consensus sur des méthodes de mesure robustes et vérifiables du risque et de la fiabilité, et sur leur applicabilité à différents cas d'utilisation de l'IA, constitue un défi pour la mesure du risque lié à l'IA. Parmi les pièges potentiels liés à la mesure des risques ou des préjudices négatifs, on peut citer le fait que l'élaboration de mesures est souvent une entreprise institutionnelle et qu'elle peut refléter par inadvertance des facteurs sans rapport avec l'impact sous-jacent. En outre, les méthodes de mesure peuvent être simplifiées à l'extrême, faire l'objet de jeux, manquer de nuances essentielles, être utilisées de manière inattendue ou ne pas tenir compte des différences entre les groupes et les contextes concernés.

Les approches visant à mesurer l'impact sur une population sont plus efficaces si elles tiennent compte du fait que les contextes sont importants, que les préjudices peuvent affecter différemment divers groupes ou sous-groupes et que les communautés ou autres sous-groupes susceptibles de subir des préjudices ne sont pas toujours des utilisateurs directs d'un système.

Risque à différents stades du cycle de vie de l'IA : La mesure des risques à un stade antérieur du cycle de vie de l'IA peut donner des résultats différents de la mesure des risques à un stade ultérieur ; certains risques peuvent être latents à un moment donné et peuvent augmenter à mesure que les systèmes d'IA s'adaptent et évoluent. En outre, les différents acteurs de l'IA tout au long du cycle de vie de l'IA peuvent avoir des perspectives différentes en matière de risque. Par , un développeur d'IA qui met à disposition des logiciels d'IA, tels que des modèles pré-entraînés, peut avoir une perspective de risque différente de celle d'un acteur de l'IA chargé de déployer ce modèle pré-entraîné dans un cas d'utilisation spécifique. Ces utilisateurs peuvent ne pas reconnaître que leur utilisation particulière peut comporter des risques différents de ceux perçus par le développeur initial. Tous les acteurs de l'IA concernés partagent la responsabilité de la conception, du développement et du déploiement d'un système d'IA digne de confiance et adapté à l'usage auquel il est destiné.

Les risques dans le monde réel : Si la mesure des risques liés à l'IA en laboratoire ou dans un environnement contrôlé peut fournir des informations importantes avant le déploiement, ces mesures peuvent différer des risques qui apparaissent dans des environnements opérationnels réels.

Inscrutabilité : Les systèmes d'IA inscrutables peuvent compliquer la mesure des risques. L'inscrutabilité peut résulter de la nature opaque des systèmes d'IA (explicabilité ou interprétabilité limitée), du manque de transparence ou de documentation dans le développement ou le déploiement des systèmes d'IA, ou des incertitudes inhérentes aux systèmes d'IA.

Base humaine : La gestion des risques liés aux systèmes d'IA destinés à augmenter ou à remplacer l'activité humaine, par exemple la prise de décision, nécessite une certaine forme de mesures de référence à des fins de comparaison. Il est difficile de systématiser cet aspect, car les systèmes d'IA exécutent des tâches différentes - et les exécutent différemment - de celles des humains.

1.2.2 Tolérance au risque

Si le cadre de référence de l'IA peut être utilisé pour hiérarchiser les risques, il ne prescrit pas la tolérance au risque. La *tolérance au risque* fait référence à la volonté de l'organisation ou de l'acteur de l'IA (voir annexe A) de supporter risque afin d'atteindre ses objectifs. La tolérance au risque peut être influencée par des exigences légales ou réglementaires (Adapté de : ISO GUIDE 73). La tolérance au risque et le niveau de risque acceptable pour les organisations ou la société sont fortement contextuels et spécifiques à l'application et au cas d'utilisation. Les tolérances au risque peuvent être influencées par les politiques et les normes établies par les propriétaires de systèmes d'IA, les organisations, les industries, les communautés ou les politiques. Les tolérances au risque sont susceptibles de changer au fil du temps, à mesure que les systèmes d'IA, les politiques et les normes évoluent. Différentes organisations peuvent avoir des tolérances au risque différentes en raison de leurs priorités organisationnelles particulières et de leurs considérations en matière de ressources.

Les entreprises, les gouvernements, les universités et la société civile continueront à développer et à débattre des connaissances et des méthodes émergentes pour mieux informer sur les compromis entre dommages et coûts-avantages. Dans la mesure où les problèmes liés à la définition des seuils de tolérance aux risques de l'IA ne sont pas résolus, il se peut que, dans certains contextes, un cadre de gestion des risques ne soit pas encore facilement applicable à l'atténuation des risques négatifs de l'IA.

Le cadre est conçu pour être flexible et pour compléter les pratiques existantes en matière de risques, qui devraient être alignées sur les lois, les règlements et les normes en vigueur. Les organisations devraient suivre les réglementations et lignes directrices existantes en matière de critères de risque, de tolérance et de réponse établies par les exigences de l'organisation, du domaine, de la discipline, du secteur ou de la profession. Certains secteurs ou industries peuvent avoir établi des définitions du préjudice ou des exigences en matière de documentation, de rapports et de divulgation. Au sein des secteurs, la gestion des risques peut dépendre des lignes directrices existantes pour des applications et des cas d'utilisation spécifiques. En l'absence de lignes directrices établies, les organisations doivent définir une tolérance raisonnable à l'égard du risque. Une fois la tolérance définie, ce cadre de référence de l'IA peut être utilisé pour gérer les risques et documenter les processus de gestion des risques.

1.2.3 Hiérarchisation des risques

Tenter d'éliminer totalement les risques négatifs peut s'avérer contre-productif dans la pratique, car tous les incidents et toutes les défaillances ne peuvent être éliminés. Des attentes irréalistes en matière de risque peuvent conduire les organisations à allouer des ressources d'une manière qui rend le triage des risques inefficace ou peu pratique, ou qui gaspille des ressources limitées. Une culture de gestion des risques peut aider les organisations à reconnaître que les risques liés à l'IA ne sont pas tous identiques et que les ressources peuvent être allouées de manière ciblée. Les efforts de gestion des risques réalisables établissent des lignes directrices claires pour évaluer la fiabilité de chaque système d'IA qu'une organisation développe ou déploie. Les politiques et les ressources doivent être hiérarchisées en fonction du niveau de risque évalué et de l'impact potentiel d'un système d'IA. La mesure dans laquelle un système d'IA peut être personnalisé ou adapté au contexte spécifique d'utilisation par le déployeur de l'IA peut être un facteur contributif.

Lors de l'application du cadre de gestion des risques liés à 'IA, les risques que l'organisation juge les plus élevés pour les systèmes d'IA dans un contexte d'utilisation donné doivent faire l'objet de la hiérarchisation la plus urgente et du processus de gestion des risques le plus approfondi. Lorsqu'un système d'IA présente des niveaux de risque négatif inacceptables - par exemple lorsque des effets négatifs importants sont imminents, que des préjudices graves se produisent réellement ou que des risques catastrophiques sont présents - le développement et le déploiement doivent cesser en toute sécurité jusqu'à ce que les risques puissent être gérés de manière satisfaisante. Si le développement, le déploiement et les cas d'utilisation d'un système d'IA s'avèrent être à faible risque dans un contexte spécifique, cela peut suggérer une priorité potentiellement moins élevée.

La hiérarchisation des risques peut différer entre les systèmes d'IA conçus ou déployés pour interagir directement avec les humains et les systèmes d'IA qui ne le sont pas. Une hiérarchisation initiale plus élevée peut s'avérer nécessaire lorsque le système d'IA est entraîné sur de grands ensembles de données comprenant des données sensibles ou protégées, telles que des informations personnelles identifiables, ou lorsque les résultats des systèmes d'IA ont un impact direct ou indirect sur les êtres humains. Les systèmes d'IA conçus pour interagir uniquement avec des systèmes informatiques et formés sur des ensembles de données non sensibles (par exemple, des données collectées dans l'environnement physique) peuvent nécessiter une priorité initiale moins élevée. Néanmoins, il est important d'évaluer et de hiérarchiser régulièrement les risques en fonction du contexte, car les systèmes d'IA qui ne sont pas en contact avec l'homme peuvent avoir des répercussions en aval sur la sécurité ou la société.

Le risque résiduel - défini comme le risque restant après le traitement du risque (Source : ISO GUIDE 73) - a un impact direct sur les utilisateurs finaux ou les personnes et communautés concernées. En documentant les risques résiduels, le fournisseur du système devra prendre pleinement en compte les risques liés au déploiement du produit d'IA et informera les utilisateurs finaux des effets négatifs potentiels de l'interaction avec le système.

1.2.4 Intégration organisationnelle et gestion des risques

Les risques liés à l'IA ne doivent pas être considérés isolément. Les différents acteurs de l'IA ont des responsabilités et une sensibilisation différentes en fonction de leur rôle dans le cycle de vie. Par exemple, les organisations qui développent un système d'IA ne disposent souvent pas d'informations sur la manière dont le système peut être utilisé. La gestion des risques liés à l'IA doit être intégrée et incorporée dans les stratégies et processus plus larges de gestion des risques de l'entreprise. Traiter les risques liés à l'IA en même temps que d'autres risques critiques, tels que la cybersécurité et la protection de la vie privée, permettra d'obtenir des résultats plus intégrés et des gains d'efficacité pour l'organisation.

Le cadre de référence pour l'IA peut être utilisé avec des orientations et des cadres connexes pour gérer les risques liés aux systèmes d'IA ou les risques plus généraux de l'entreprise. Certains risques liés aux systèmes d'IA sont communs à d'autres types de développement et de déploiement de logiciels. Parmi les exemples de risques qui se recoupent, on peut citer : les problèmes de protection de la vie privée liés à l'utilisation de données sous-jacentes pour entraîner les systèmes d'IA ; les implications énergétiques et environnementales associées aux demandes de calcul à forte intensité de ressources ; les problèmes de sécurité liés à la confidentialité, à l'intégrité et à la disponibilité du système et de ses données d'entraînement et de sortie ; et la sécurité générale du logiciel et du matériel sous-jacents pour les systèmes d'IA.

Pour que la gestion des risques soit efficace, les organisations doivent mettre en place et maintenir des mécanismes de responsabilité, des rôles et des responsabilités, une culture et des structures d'incitation appropriés. L'utilisation du cadre de référence de l'IA ne suffira pas à entraîner ces changements ou à fournir les incitations appropriées. Une gestion efficace des risques passe par un engagement organisationnel au plus haut niveau et peut nécessiter un changement culturel au sein d'une organisation ou d'un secteur d'activité. En outre, les petites et moyennes organisations qui gèrent les risques liés à l'IA ou qui mettent en œuvre le cadre de référence pour l'IA peuvent être confrontées à des défis différents de ceux des grandes organisations, en fonction de leurs capacités et de leurs ressources.

2. Public

L'identification et la gestion des risques liés à l'IA et de ses incidences potentielles - tant positives que négatives - requièrent un large éventail de perspectives et d'acteurs tout au long du cycle de vie de l'IA. Idéalement, acteurs de l'IA représenteront une diversité d'expériences, d'expertises et de formations et constitueront des équipes diversifiées sur le plan démographique et disciplinaire. Le cadre de référence de l'IA est destiné à être utilisé par les acteurs de l'IA tout au long du cycle de vie et des dimensions de l'IA.

L'OCDE a élaboré un cadre de classification des activités du cycle de vie de l'IA en fonction de cinq dimensions sociotechniques clés, chacune ayant des propriétés pertinentes pour la politique et la gouvernance de l'IA, y compris la gestion des risques [OCDE (2022) OECD Framework for the Classification of AI systems - OECD Digital Economy Papers]. La figure 2 illustre ces dimensions, légèrement modifiées par le NIST aux fins du présent cadre. La modification apportée par le NIST souligne l'importance des processus de test, d'évaluation, de vérification et de validation (TEVV) tout au long du cycle de vie de l'IA et généralise le contexte opérationnel d'un système d'IA.

Les dimensions de l'IA présentées dans la figure 2 sont le contexte de l'application, les données et les entrées, le modèle d'IA et les tâches et sorties. Les acteurs de l'IA impliqués dans ces dimensions qui effectuent ou gèrent la conception, le développement, le déploiement, l'évaluation et l'utilisation des systèmes d'IA et qui dirigent les efforts de gestion des risques liés à l'IA constituent le *principal* public du CMR sur l'IA.

Les acteurs représentatifs de l'IA dans les différentes dimensions du cycle de vie sont énumérés à la figure 3 et décrits en détail à l'annexe A. Dans le cadre du RMF de l'IA, tous les acteurs de l'IA travaillent ensemble pour gérer les risques et atteindre les objectifs d'une IA digne de confiance et responsable. Les acteurs de l'IA disposant d'une expertise spécifique en matière de TEVV sont intégrés tout au long du cycle de vie de l'IA et sont particulièrement susceptibles de bénéficier du cadre. Exécutées régulièrement, les tâches TEVV peuvent fournir des informations relatives aux techniques, sociétales, juridiques et éthiques, et peuvent aider à anticiper les impacts ainsi qu'à évaluer et à suivre les risques émergents. En tant que processus régulier dans le cadre du cycle de vie de l'IA, la TEVV permet à la fois de prendre des mesures correctives à mi-parcours et de gérer les risques a posteriori.

La dimension "People & Planet", au centre de la figure 2, représente les droits de l'homme et le bien-être général de la société et de la planète. Les acteurs de l'IA dans cette dimension constituent un public distinct du CMR de l'IA qui *informe* le public principal. Ces acteurs de l'IA peuvent être des associations commerciales, des organismes de normalisation, des chercheurs, des groupes de défense,

Fig. 2. Cycle de vie et dimensions clés d'un système d'IA. Modifié d'après OCDE (2022) Cadre de l'pour la classification des systèmes d'IA - Documents de l'OCDE sur l'économie numériqueOCDE . Les deux cercles intérieurs représentent les dimensions clés des systèmes d'IA et le cercle extérieur les étapes du cycle de vie de l'IA. Idéalement, les efforts de gestion des risques commencent par la fonction de planification et de conception dans le contexte de l'application et se poursuivent tout au long du cycle de vie du système d'IA. La figure 3 présente des acteurs représentatifs de l'IA.

les groupes environnementaux, les organisations de la société civile, les utilisateurs finaux et les personnes et communautés potentiellement touchées. Ces acteurs peuvent :

- aider à fournir un contexte et à comprendre les impacts potentiels et réels ;
- être une source de normes et d'orientations formelles ou quasi-formelles pour la gestion des risques liés à l'IA ;
- définir les limites du fonctionnement de l'IA (techniques, sociétales, juridiques et éthiques) ; et
- promouvoir la discussion sur les compromis nécessaires pour équilibrer les valeurs et les priorités sociétales liées aux libertés et aux droits civils, à l'équité, à l'environnement et à la planète, ainsi qu'à l'économie.

La réussite de la gestion des risques dépend du sens de la responsabilité collective des acteurs de l'IA, comme le montre la figure 3. Les fonctions du CMR de l'IA, décrites à la section 5, requièrent des perspectives, des disciplines, des professions et des expériences diverses. La diversité des équipes contribue à un partage plus ouvert des idées et des hypothèses sur les objectifs et les fonctions de la technologie, rendant ainsi ces aspects implicites plus explicites. Cette perspective collective élargie permet de mettre en évidence les problèmes et d'identifier les risques existants et émergents.

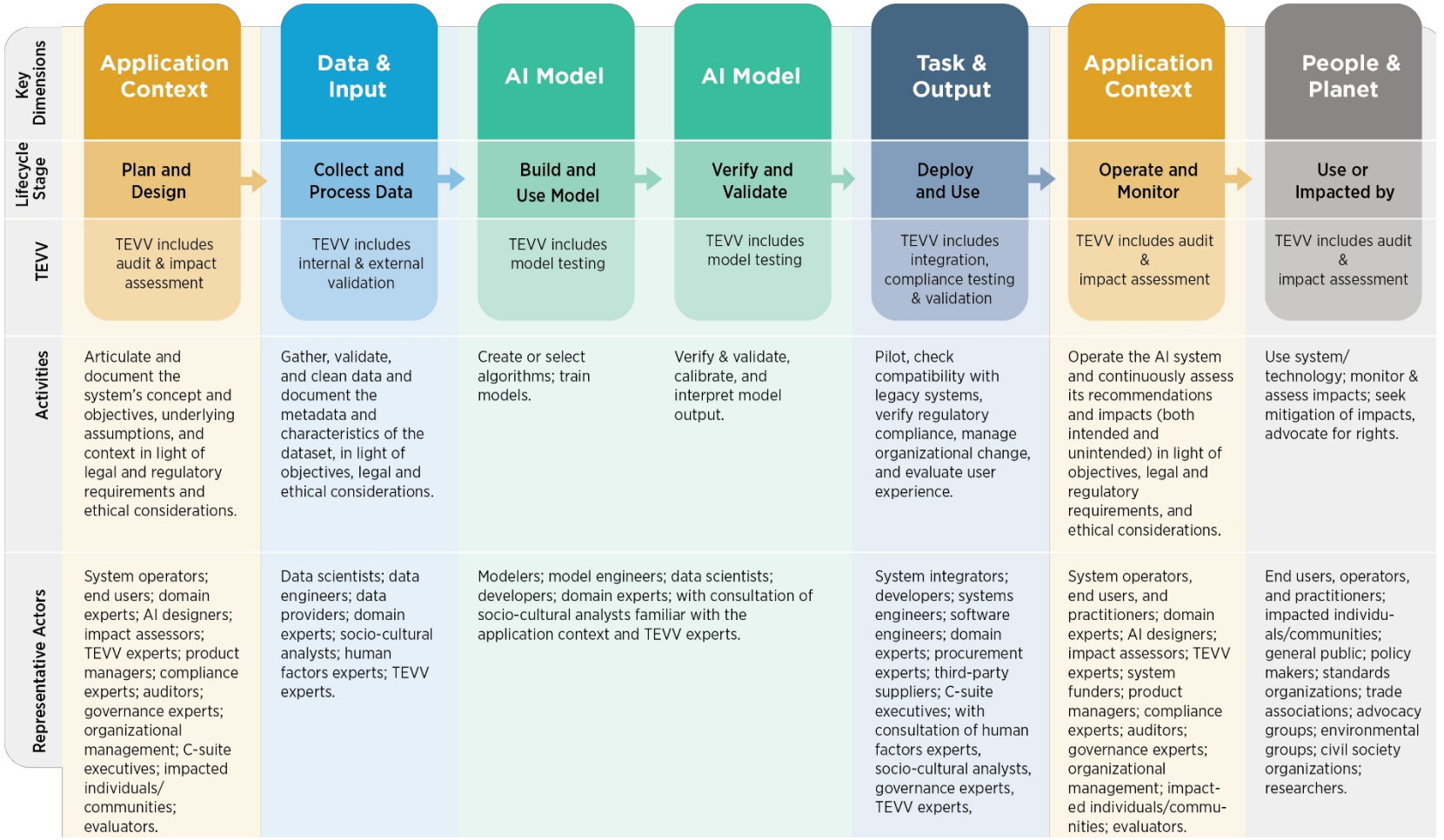

Fig. 3. Acteurs de l'IA aux différentes étapes du cycle de vie de l'IA. Voir l'annexe A pour des descriptions détaillées des tâches des acteurs de l'IA, y compris des détails sur les tâches de test, d'évaluation, de vérification et de validation. Il convient de noter que les acteurs de l'IA dans la dimension du modèle d'IA (figure 2) sont séparés en tant que meilleure pratiqueceux qui construisent et utilisent les modèles étant séparés de ceux qui vérifient et valident les modèles.

3. Risques liés à l'IA et fiabilité

Pour que les systèmes d'IA soient dignes de confiance, ils doivent souvent répondre à une multiplicité de cri- tères qui ont de la valeur pour les parties intéressées. Les approches qui renforcent la fiabilité de l'IA peuvent réduire les risques négatifs liés à l'IA. Le présent cadre définit **les caractéristiques** suivantes de l'IA digne de confiance et propose des orientations pour répondre. Les caractéristiques des systèmes d'IA dignes de confiance sont les suivantes : **valides et fiables, sûrs, sécurisés et résilients, responsables et trans-parents, explicables et interprétables, respectueux de la vie privée, et équitables avec gestion des biais nuisibles.** Pour créer une IA digne de confiance, il faut équilibrer chacune de ces caractéristiques en fonction du contexte d'utilisation du système d'IA. Bien que toutes les caractéristiques soient des attributs du système sociotechnique, la responsabilité et la transparence concernent également les processus et les activités internes d'un système d'IA et son environnement externe. Négliger ces caractéristiques peut augmenter la probabilité et l'ampleur des conséquences négatives.

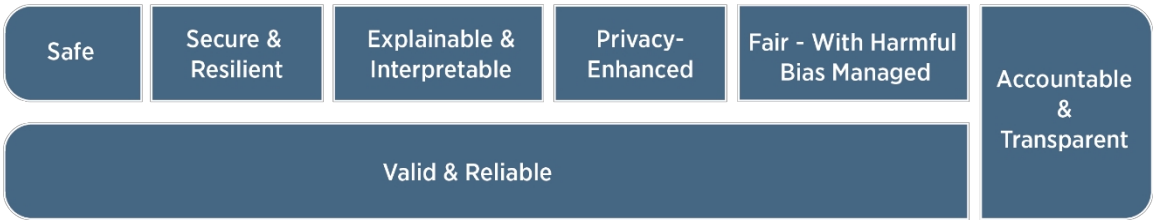

Fig. 4. Caractéristiques des systèmes d'IA dignes de confiance. Valide et fiable est une condition nécessaire à la fiabilité et est présentée comme la base des autres caractéristiques de fiabilité. La notion de responsabilité et de transparence est représentée par une case verticale parce qu'elle est liée à toutes les autres caractéristiques.

Les caractéristiques de fiabilité (illustrées à la figure 4) sont inextricablement liées au comportement social et organisationnel, aux ensembles de données utilisés par les systèmes d'IA, à la sélection des modèles et algorithmes d'IA et aux décisions prises par ceux qui construisent, ainsi qu'aux interactions avec les êtres humains qui fournissent des informations à partir de ces systèmes et les supervisent. Il convient de faire appel au jugement humain pour décider des paramètres spécifiques liés aux caractéristiques de fiabilité de l'IA et valeurs seuils précises pour ces paramètres.

Le fait d'aborder individuellement les caractéristiques de la fiabilité de l'IA ne garantira pas la fiabilité du système d'IA ; des compromis sont généralement nécessaires, il est rare que toutes les caractéristiques s'appliquent dans tous les cas, et certaines seront plus ou moins importantes dans une situation donnée. En fin de compte, la confiance est un concept social qui s'étend sur tout un spectre et qui n'est aussi fort que ses caractéristiques les plus faibles.

Lors de la gestion des risques liés à l'IA, les organisations peuvent être confrontées à des décisions difficiles pour équilibrer ces caractéristiques. Par , dans certains scénarios, des compromis peuvent apparaître entre l'optimisation de l'interprétabilité et la protection de la vie privée. Dans d'autres cas, les organisations peuvent être confrontées à un compromis entre la précision prédictive et l'interprétabilité. Ou encore, dans certaines conditions telles que la rareté des données, les techniques d'amélioration de la protection de la vie privée peuvent entraîner une perte de précision, ce qui affecte les décisions.

sur l'équité et d'autres valeurs dans certains domaines. Pour traiter les compromis, il faut tenir compte du contexte décisionnel. Ces analyses peuvent mettre en évidence l'existence et l'étendue des compromis entre différentes mesures, mais elles ne répondent pas aux questions relatives à la manière de gérer ces compromis. Celles-ci dépendent des valeurs en jeu dans le *contexte* concerné et doivent être résolues d'une manière à la fois transparente et justifiable.

Il existe de multiples approches pour améliorer la connaissance du contexte dans le cycle de vie de l'IA. Par exemple, les experts en la matière peuvent contribuer à l'évaluation des résultats de la TEVV et collaborer avec les équipes chargées des produits et du déploiement afin d'aligner les paramètres de la TEVV sur les exigences et les conditions de déploiement. Lorsque les ressources nécessaires sont disponibles, l'augmentation de l'ampleur et de la diversité des contributions des parties intéressées et des acteurs concernés par l'IA tout au long du cycle de vie de l'IA peut améliorer les possibilités d'informer les évaluations sensibles au contexte et d'identifier les avantages et les incidences positives des systèmes d'IA. Ces pratiques peuvent accroître la probabilité que les risques découlant des contextes sociaux soient gérés de manière appropriée.

La compréhension et le traitement des caractéristiques de fiabilité dépendent du rôle particulier de l'acteur de l'IA dans le cycle de vie de l'IA. Pour un système d'IA donné, un concepteur ou un développeur d'IA peut avoir une perception des caractéristiques différente de celle du déployeur.

Les caractéristiques de fiabilité expliquées dans le présent document s'influencent mutuellement. Les systèmes hautement sécurisés mais injustes, les systèmes précis mais opaques et ininterprétables, et les systèmes imprécis mais sécurisés, respectueux de la vie privée et transparents sont tous indésirables. Une approche globale de la gestion des risques exige de trouver un entre les caractéristiques de fiabilité. Il incombe à tous les acteurs de l'IA de déterminer si la technologie de l'IA est un outil approprié ou nécessaire pour un contexte ou un objectif donné, et comment l'utiliser de manière responsable. La décision de mettre en service ou de déployer un système d'IA devrait être fondée sur une évaluation contextuelle des caractéristiques de fiabilité et des risques, incidences, coûts et avantages relatifs, et s'appuyer sur un large éventail de parties intéressées.

3.1 Valable et fiable

La validation est la "confirmation, par la fourniture de preuves objectives, que les conditions requises pour une utilisation ou une application spécifique prévue ont été remplies" (Source : ISO 9000:2015). Le déploiement de systèmes d'IA inexacts, peu fiables ou mal généralisés à des données et à des contextes dépassant le cadre de leur formation crée et augmente les risques négatifs liés à l'IA et réduit la fiabilité.

La *fiabilité* est définie dans la même norme comme "l'aptitude d'un élément à fonctionner comme requis, sans défaillance, pendant un intervalle de temps donné, dans des conditions données" (Source : ISO/IEC TS 5723:2022). La fiabilité est un objectif de correction globale du fonctionnement d'un système d'IA dans les conditions d'utilisation prévues et sur une donnée, y compris pendant toute la durée de vie du système.

L'exactitude et la robustesse contribuent à la validité et à la fiabilité des systèmes d'IA et peuvent être en conflit dans les systèmes d'IA.

La *précision* est définie par la norme ISO/IEC TS 5723:2022 comme "la proximité des résultats d'observations, de calculs ou d'estimations par rapport aux valeurs vraies ou aux valeurs acceptées comme étant vraies". Les mesures de la précision devraient tenir compte des mesures centrées sur le calcul (par exemple, les taux de faux positifs et de faux négatifs), de la collaboration entre l'homme et l'IA, et démontrer une validité externe (généralisable au-delà des conditions de formation). Les mesures de précision doivent toujours être associées à des ensembles d'essais clairement définis et réalistes - qui sont représentatifs des conditions d'utilisation prévues - et à des détails sur la méthodologie d'essai ; ces éléments doivent figurer dans la documentation associée. Les mesures de précision peuvent inclure une désagrégation des résultats pour différents segments de données.

La robustesse ou la *généralisabilité* est définie comme "l'aptitude d'un système à maintenir niveau de performance dans diverses circonstances" (Source : ISO/IEC TS 5723:2022). La robustesse est l'objectif d'une fonctionnalité appropriée du système dans un large éventail de conditions et de circonstances, y compris des utilisations des systèmes d'IA non prévues au départ. La robustesse exige non seulement que le système fonctionne exactement comme il le fait dans les utilisations prévues, mais aussi qu'il fonctionne de manière à minimiser les dommages potentiels pour les personnes s'il fonctionne dans un contexte inattendu.

La validité et la fiabilité des systèmes d'IA déployés sont souvent évaluées par des tests ou des contrôles continus qui confirment qu'un système fonctionne comme prévu. La mesure de la validité, de l'exactitude, de la robustesse et de la fiabilité contribue à la fiabilité et devrait tenir du fait que certains types de défaillances peuvent causer des dommages plus importants. Les efforts de gestion des risques liés à l'IA devraient viser en priorité à minimiser les incidences négatives potentielles et pourraient nécessiter une intervention humaine dans les cas où le système d'IA ne peut pas détecter ou corriger les erreurs.

3.2 Sûr

Les systèmes d'IA ne doivent pas "dans des conditions définies, conduire à un état dans lequel la vie humaine, la santé, les biens ou l'environnement sont mis en danger" (Source : ISO/IEC TS 5723:2022). La sécurité de fonctionnement des systèmes d'IA est améliorée par

- des pratiques responsables en matière de conception, de développement et de déploiement ;
- des informations claires aux déployeurs sur l'utilisation responsable du système ;
- une prise de décision responsable de la part des déployeurs et des utilisateurs finaux ; et
- explications et documentation des risques sur la base de preuves empiriques d'incidents.

Les différents types de risques de sécurité peuvent nécessiter des approches de gestion des risques d'IA adaptées en fonction du contexte et de la gravité des risques potentiels présentés. Les risques de sécurité qui présentent un risque potentiel de blessure grave ou de décès requièrent la hiérarchisation la plus urgente et le processus de gestion des risques le plus approfondi.

La prise en compte de la sécurité tout au long du cycle de vie et dès le début de la planification et de la conception permet d'éviter les défaillances ou les conditions qui peuvent rendre un système dangereux. D'autres approches pratiques de la sécurité de l'IA sont souvent liées à une simulation rigoureuse et à des essais sur le terrain, à une surveillance en temps réel et à la possibilité d'arrêter, de modifier ou de faire intervenir l'homme dans les systèmes qui s'écartent de la fonctionnalité prévue ou attendue.

Les approches de gestion des risques liés à la sécurité de l'IA devraient s'inspirer des efforts et des lignes directrices en matière de sécurité dans des domaines tels que les transports et les soins de santé, et s'aligner sur les lignes directrices ou les normes existantes spécifiques à un secteur ou à une application.

3.3 Sécurité et résilience

Les systèmes d'IA, ainsi que les écosystèmes dans lesquels ils sont déployés, peuvent être considérés comme *résilients* s'ils peuvent résister à des événements défavorables inattendus ou à des changements imprévus dans leur environnement ou leur utilisation - ou s'ils peuvent maintenir leurs fonctions et leur structure face à des changements internes et externes et se dégrader de manière sûre et gracieuse lorsque cela est nécessaire (adapté de : ISO/IEC TS 5723:2022). Les problèmes de sécurité les plus courants concernent les exemples adverses, l'empoisonnement des données et l'exfiltration de modèles, de données d'entraînement ou d'autres propriétés intellectuelles par le biais des points d'extrémité des systèmes d'IA. Les systèmes d'IA qui peuvent maintenir la confidentialité, l'intégrité et la disponibilité grâce à des mécanismes de protection qui empêchent l'accès et l'utilisation non autorisés peuvent être considérés comme *sûrs*. Les lignes directrices du cadre de cybersécurité du NIST et du cadre de gestion des risques sont notamment applicables ici.

La sécurité et la résilience sont des caractéristiques liées mais distinctes. Alors que la résilience est la capacité de revenir à un fonctionnement normal après un événement négatif inattendu, la sécurité inclut la résilience mais aussi les protocoles visant à éviter les attaques, à s'en protéger, à y répondre ou à s'en remettre. La résilience est liée à la robustesse et va au-delà de la provenance des données pour englober l'utilisation inattendue ou malveillante (ou l'abus ou la mauvaise utilisation) du modèle ou des données.

3.4 Responsabilité et transparence

L'IA digne de confiance dépend de la responsabilité. La responsabilité présuppose la transparence. *La transparence* reflète la mesure dans laquelle les informations relatives à un système d'IA et à ses résultats sont accessibles aux personnes qui interagissent avec ce système, qu'elles soient conscientes ou non. Une transparence significative permet d'accéder à des niveaux d'information appropriés en fonction du stade du cycle de vie de l'IA et adaptés au rôle ou aux connaissances des acteurs de l'IA ou des personnes qui interagissent avec le système d'IA ou qui l'utilisent. En favorisant des niveaux de compréhension plus élevés, la transparence accroît la confiance dans le système d'IA.

Le champ d'application de cette caractéristique s'étend des décisions de conception et des données de formation à la formation au modèle, à la structure du modèle, aux cas d'utilisation prévus et à la manière dont les décisions relatives au déploiement, au post-déploiement ou à l'utilisateur final ont été prises et à quel moment, et par qui. La transparence est souvent nécessaire pour obtenir des réparations en cas de résultats incorrects ou d'impacts négatifs des systèmes d'IA. La transparence doit tenir compte de l'interaction entre l'homme et l'IA : par exemple

la manière dont un opérateur ou un utilisateur humain est informé de la détection d'un résultat négatif potentiel ou réel causé par un système d'IA. Un système transparent n'est pas nécessairement un système précis, respectueux de la vie privée, sûr ou équitable. Toutefois, il est difficile déterminer si un système opaque possède de telles caractéristiques, et de le faire au fil du temps, à mesure que les systèmes complexes évoluent.

Le rôle des acteurs de l'IA doit être pris en compte lorsqu'il s'agit de rendre compte des résultats systèmes d'IA. La relation entre le risque et la responsabilité associée à l'IA et aux systèmes technologiques en général diffère selon les contextes culturels, juridiques, sectoriels et sociétaux. Lorsque les conséquences sont graves, par exemple lorsque la vie et la liberté sont en jeu, les développeurs et les utilisateurs de l'IA devraient envisager d'adapter de manière proportionnelle et proactive leurs pratiques en matière de transparence et de responsabilité. Le maintien de pratiques organisationnelles et de structures de gouvernance pour la réduction des risques, comme la gestion des risques, peut contribuer à la mise en place de systèmes plus responsables.

Les mesures visant à renforcer la transparence et la responsabilité doivent également tenir compte de l'impact de ces efforts sur l'entité chargée de la mise en œuvre, y compris le niveau des ressources nécessaires et nécessité de protéger les informations confidentielles.

Le maintien de la provenance des données d'entraînement et l'attribution des décisions du système d'IA à des sous-ensembles de données d'entraînement peuvent contribuer à la fois à la transparence et à la responsabilité. Les données d'entraînement peuvent également être soumises à des droits d'auteur et doivent respecter les lois applicables en matière de droits de propriété intellectuelle.

Comme les outils de transparence pour les systèmes d'IA et la documentation connexe continuent d'évoluer, les développeurs de systèmes d'IA sont encouragés à tester différents types d'outils de transparence en coopération avec les déployeurs d'IA pour s'assurer que les systèmes d'IA sont utilisés comme prévu.

3.5 Explicable et interprétable

L'explicabilité fait référence à une représentation des mécanismes qui sous-tendent le fonctionnement des systèmes d'IA, tandis que l'*interprétabilité* fait référence à la signification des résultats des systèmes d'IA dans le contexte des objectifs fonctionnels pour lesquels ils ont été conçus. Ensemble, l'explicabilité et l'interprétabilité aident ceux qui exploitent ou supervisent un système d'IA, ainsi que les utilisateurs de ce système, à mieux comprendre la fonctionnalité et la fiabilité du système, y compris ses résultats. L'hypothèse sous-jacente est que la perception d'un risque négatif découle d'un manque de capacité à donner un sens ou à contextualiser les résultats du système de manière appropriée. Les systèmes d'IA explicables et interprétables fournissent des informations qui aideront les utilisateurs finaux à comprendre les objectifs et l'impact potentiel d'un système d'IA.

Le risque lié au manque d'explicabilité peut être géré en décrivant le fonctionnement des systèmes d'IA, avec des descriptions adaptées aux différences individuelles telles que le rôle, les connaissances et le niveau de compétence de l'utilisateur. Les systèmes explicables peuvent être débogués et contrôlés plus facilement et se prêtent à une documentation, un audit et une gouvernance plus approfondis.

Les risques liés à l'interprétabilité peuvent souvent être résolus en communiquant une description des raisons pour lesquelles un système d'intelligence artificielle a fait une prédiction ou une recommandation particulière. (Voir "Quatre principes de l'intelligence artificielle explicable" et "Fondements psychologiques de l'explicabilité et de l'interprétabilité dans l'intelligence artificielle" ici).

La transparence, l'explicabilité et l'interprétabilité sont des caractéristiques distinctes qui se renforcent mutuellement. La transparence permet de répondre à la question de savoir "ce qui s'est passé" dans le système. L'explicabilité peut répondre à la question de savoir "comment" une décision a été prise dans le système. L'inter-prétabilité peut répondre à la question de savoir "pourquoi" une décision a été prise par le système et quelle est sa signification ou son contexte pour l'utilisateur.

3.6 Renforcement de la protection de la vie privée

La protection de la vie privée fait généralement référence aux normes et pratiques qui contribuent à sauvegarder l'autonomie, l'identité et la dignité humaines. Ces normes et pratiques concernent généralement la liberté d'intrusion, la limitation de l'observation ou le pouvoir des individus de consentir à la divulgation ou au contrôle des facettes leur identité (par exemple, leur corps, leurs données, leur réputation). (Voir le cadre de protection de la vie privée du NIST : Un outil pour améliorer la protection de la vie privée par la gestion des risques de l'entreprise).

Les valeurs de protection de la vie privée telles que l'anonymat, la confidentialité et le contrôle devraient généralement guider les choix en matière de conception, de développement et de déploiement des systèmes d'IA. Les risques liés à la vie privée peuvent influer sur la sécurité, la partialité et la transparence et s'accompagner de compromis avec ces autres caractéristiques. À l'instar de la sûreté et de la sécurité, les caractéristiques techniques spécifiques d'un système d'IA peuvent favoriser ou réduire la protection de la vie privée. Les systèmes d'IA peuvent également présenter de nouveaux risques pour la vie privée en permettant par inférence d'identifier des personnes ou des informations précédemment privées sur des personnes.

Les technologies d'amélioration de la confidentialité ("PET") pour l'IA, ainsi que les méthodes de minimisation des données telles que la dépersonnalisation et l'agrégation pour certains résultats de modèles, peuvent contribuer à la conception de systèmes d'IA améliorés sur le plan de la confidentialité. Dans certaines conditions, telles que la rareté des données, les techniques d'amélioration de la confidentialité peuvent entraîner une perte de précision, affectant les décisions relatives à l'équité et à d'autres valeurs dans certains domaines.

3.7 Équitable - avec gestion des préjugés nuisibles

L'équité dans l'IA englobe les préoccupations en matière d'égalité et d'équité en abordant des questions telles que les préjugés et les discriminations préjudiciables. Les normes d'équité peuvent être complexes et difficiles à définir, car les perceptions de l'équité diffèrent d'une culture à l'autre et peuvent évoluer en fonction de l'application. Les efforts de gestion des risques des organisations seront renforcés par la reconnaissance et la prise en compte de ces différences. Les systèmes dans lesquels les biais nuisibles sont atténués ne sont pas nécessairement équitables. Par exemple, les systèmes dans lesquels les prévisions sont quelque peu équilibrées entre les groupes démographiques peuvent encore être inaccessibles aux personnes handicapées ou touchées par la fracture numérique, ou peuvent exacerber les disparités existantes ou les préjugés systémiques.

Les biais vont au-delà de l'équilibre démographique et de la représentativité des données. Le NIST a identifié trois grandes catégories de biais d'IA à prendre en compte et à gérer : systémique, informatique et statistique, et humain-cognitif. Chacune de ces catégories peut se produire en l'absence de préjugés, de partialité ou d'intention discriminatoire. Les biais systémiques peuvent être présents dans les ensembles de données d'IA, les normes, pratiques et processus organisationnels tout au long du cycle de vie de l'IA, et la société au sens large qui utilise les systèmes d'IA. Les biais informatiques et statistiques peuvent être présents dans les ensembles de données d'IA et les processus algorithmiques, et découlent souvent d'erreurs systématiques dues à des échantillons non représentatifs. Les biais cognitifs humains sont liés à la manière dont un individu ou un groupe perçoit les informations du système d'IA pour prendre une décision ou compléter des informations manquantes, ou à la manière dont les humains pensent aux objectifs et aux fonctions d'un système d'IA. Les biais cognitifs humains sont omniprésents dans les processus décisionnels tout au long du cycle de vie de l'IA et de l'utilisation du système, y compris la conception, la mise en œuvre, l'exploitation et la maintenance de l'IA.

Les préjugés existent sous de nombreuses formes et peuvent s'enraciner dans les systèmes automatisés qui nous aident à prendre des décisions concernant notre vie. Bien que les préjugés ne soient pas toujours un phénomène négatif, les systèmes d'IA peuvent potentiellement augmenter la vitesse et l'ampleur des préjugés et perpétuer et amplifier les préjudices subis par les individus, les groupes, les communautés, les organisations et la société. Les préjugés sont étroitement associés aux concepts de transparence et d'équité dans la société. (Pour plus d'informations sur les biais, y compris les trois catégories, voir la publication spéciale 1270 du NIST, Towards a Standard for Identifying and Managing Bias in Artificial Intelligence).

4. Efficacité du FER de l'IA

Les évaluations de l'efficacité des RMF d'IA - y compris les moyens de mesurer l'amélioration de la fiabilité des systèmes d'IA - feront partie des activités futures du NIST, en collaboration avec la communauté de l'IA.

Les organisations et autres utilisateurs du cadre sont encouragés à évaluer périodiquement si le cadre de gestion des risques liés à l'IA a amélioré leur capacité à gérer les risques liés à l'IA, y compris, mais sans s'y limiter, leurs politiques, processus, pratiques, plans de mise en œuvre, indicateurs, mesures et résultats escomptés. Le NIST a l'intention de travailler en collaboration avec d'autres pour développer des métriques, des méthodologies et des objectifs afin d'évaluer l'efficacité du cadre de gestion des risques liés à l'IA, et de partager largement les résultats et les informations à l'appui. Les utilisateurs du cadre devraient bénéficier de ce qui suit :

- des processus améliorés pour gouverner, cartographier, mesurer et gérer les risques liés à l'IA, et documenter clairement les résultats ;
- une meilleure prise de conscience des relations et des compromis entre les caractéristiques de confiance, les approches sociotechniques et les risques liés à l'IA ;
- des processus explicites pour prendre les décisions de mise en service et de déploiement du système ;
- a établi des politiques, des processus, des pratiques et des procédures visant à améliorer les efforts de responsabilisation de l'organisation en ce qui concerne les risques liés au système d'IA ;
- une culture organisationnelle renforcée qui donne la priorité à l'identification et à la gestion des risques liés au système d'IA et des incidences potentielles sur les individus, les communautés, les organisations et la société ;
- un meilleur partage de l'information au sein des organisations et entre elles sur les risques, les processus de prise de décision, les responsabilités, les pièges courants, les pratiques TEVV et les approches d'amélioration continue ;
- une meilleure connaissance du contexte pour une meilleure prise de conscience des risques en aval ;
- renforcer l'engagement avec les parties intéressées et les acteurs de l'IA concernés ; et
- une capacité accrue de VTEP des systèmes d'IA et des risques associés.

Partie 2 : Noyau et profils

5. AI RMF Core

Le noyau du RMF sur l'IA fournit des résultats et des actions qui favorisent le dialogue, la compréhension et les activités visant à gérer les risques liés à l'IA et à développer de manière responsable des systèmes d'IA dignes de confiance. Comme l'illustre la figure 5, le noyau est composé de quatre fonctions : **GOUVERNER, CARTOGRAPHIER, MESURER** et **GÉRER**. Chacune de ces fonctions de haut niveau est divisée en catégories et sous-catégories. Les catégories et sous-catégories sont subdivisées en actions et résultats spécifiques. Les actions ne constituent pas une liste de contrôle et ne sont pas nécessairement un ensemble ordonné d'étapes.

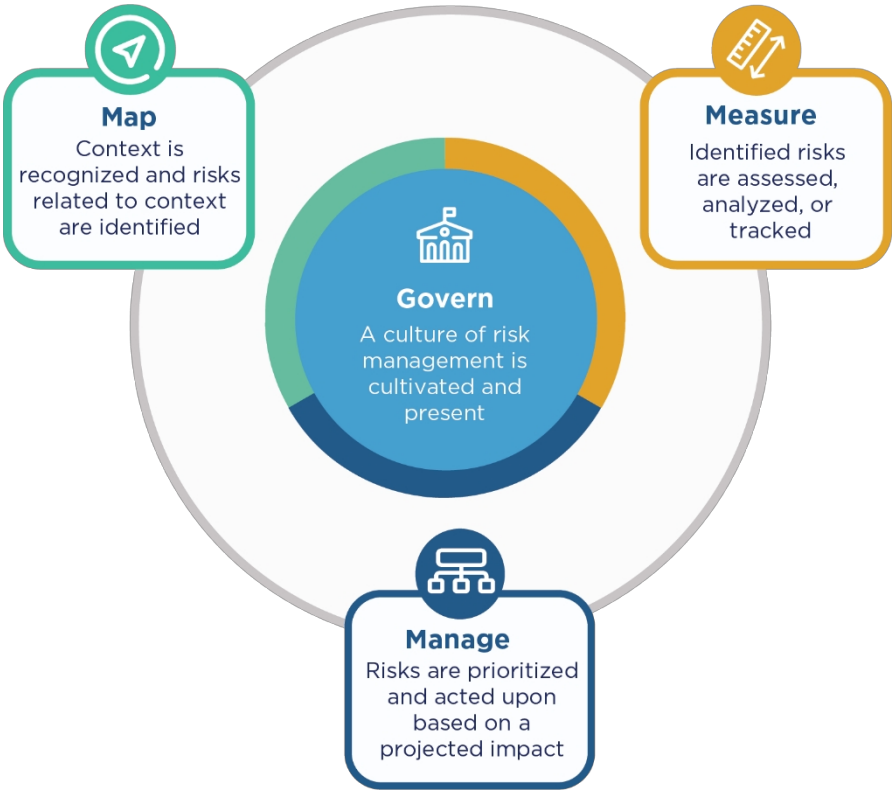

Fig. 5. Les fonctions organisent les activités de gestion des risques liés à l'IA à leur niveau le plus élevé afin de gouverner, de cartographier, de mesurer et de gérer les risques liés à l'IA. La gouvernance est conçue comme une fonction transversale qui informe et irrigue les trois autres fonctions.

La gestion des risques devrait être continue, opportune et réalisée tout au long des dimensions du cycle de vie du système d'IA. Les fonctions essentielles du CMR de l'IA devraient être exécutées de manière à refléter des perspectives diverses et pluridisciplinaires, incluant éventuellement les points de vue d'acteurs de l'IA extérieurs à l'organisation. La diversité de l'équipe contribue à un partage plus ouvert des idées et des hypothèses sur les objectifs et les fonctions de la technologie en cours de conception et de développement,

déployés ou évalués - ce qui peut créer des occasions de mettre à jour des problèmes et d'identifier des risques existants ou émergents.

Une ressource en ligne accompagnant le AI RMF, le NIST AI RMF Playbook, est disponible pour aider les organisations à naviguer dans le AI RMF et à atteindre ses résultats grâce à des suggestions d'actions tactiques qu'elles peuvent appliquer dans leur propre contexte. Tout comme l'AI RMF, le Playbook est facultatif et les organisations peuvent utiliser les suggestions en fonction de leurs besoins et de leurs intérêts. Les utilisateurs du Playbook peuvent créer des conseils personnalisés à partir des documents suggérés pour leur propre usage et faire part de leurs suggestions à l'ensemble de la communauté. Tout comme l'AI RMF, le Playbook fait partie du centre de ressources du NIST sur l'IA digne de confiance et responsable.

Les utilisateurs du cadre peuvent appliquer ces fonctions de la manière la plus adaptée à leurs besoins en matière de gestion des risques liés à l'IA, en fonction de leurs ressources et de leurs capacités. Certaines organisations peuvent choisir de sélectionner l'une des catégories et sous-catégories ; d'autres peuvent choisir et avoir la capacité d'appliquer toutes les catégories et sous-catégories. Si une structure de gouvernance est en place, les fonctions peuvent être exécutées dans n'importe quel ordre tout au long du cycle de vie de l'IA, si l'utilisateur du cadre estime qu'elles apportent une valeur ajoutée. Après avoir mis en place les résultats dans la **fonction GOUVERNER**, la plupart des utilisateurs du cadre de référence pour l'IA commenceront par la fonction **MAP** et poursuivront avec la fonction **MESURER** ou **GÉRER**. Quelle que soit la manière dont les utilisateurs intègrent les fonctions, le processus doit être itératif, avec des références croisées entre les fonctions si nécessaire. De même, certaines catégories et sous-catégories comportent des éléments qui s'appliquent à plusieurs fonctions, ou qui devraient logiquement intervenir avant certaines décisions relatives aux sous-catégories.

5.1 Gouverner

La fonction **GOVERN** :

- cultive et met en œuvre une culture de gestion des risques au sein des organisations qui conçoivent, développent, déploient, évaluent ou acquièrent des systèmes d'IA ;
- décrit les processus, les documents et les schémas organisationnels qui anticipent, identifient et gèrent les risques qu'un système peut présenter, y compris pour les utilisateurs et d'autres personnes au sein de la société, ainsi que les procédures permettant d'atteindre ces résultats ;
- intègre des processus d'évaluation des impacts potentiels ;
- fournit une structure permettant aux fonctions de gestion des risques liés à l'IA de s'aligner sur les principes, les politiques et les priorités stratégiques de l'organisation ;
- relie les aspects techniques de la conception et du développement des systèmes d'IA aux valeurs et principes de l'organisation, et permet aux personnes impliquées dans l'acquisition, la formation, le déploiement et le contrôle de ces systèmes d'acquérir des pratiques et des compétences organisationnelles ; et
- porte sur l'ensemble du cycle de vie des produits et des processus associés, y compris les questions juridiques et autres concernant l'utilisation de logiciels ou de systèmes matériels et de données de tiers.

GOVERN est une fonction transversale qui s'inscrit dans l'ensemble de la gestion des risques liés à l'IA et qui permet de mettre en œuvre les autres fonctions du processus. Les aspects de la **gouvernance**, en particulier ceux liés à la conformité ou à l'évaluation, doivent être intégrés dans chacune des autres fonctions. L'attention portée à la gouvernance est une exigence permanente et intrinsèque pour une gestion efficace des risques liés à l'IA tout au long de la durée de vie d'un système d'IA et de la hiérarchie de l'organisation.

Une gouvernance solide peut stimuler et renforcer les pratiques et les normes internes afin de faciliter la culture du risque au sein de l'organisation. Les autorités dirigeantes peuvent déterminer les politiques générales qui orientent la mission, les objectifs, les valeurs, la culture et la tolérance au risque d'une organisation. Les dirigeants donnent le ton en matière de gestion des risques au sein de l'organisation et, partant, de culture organisationnelle. La direction aligne les aspects techniques de la gestion des risques liés à l'IA sur les politiques et les opérations. La documentation peut accroître la transparence, améliorer les processus d'examen humain et renforcer la responsabilité des équipes chargées des systèmes d'IA.

Après avoir mis en place les structures, les systèmes, les processus et les équipes décrits dans la fonction **GOV- ERN**, les organisations devraient bénéficier d'une culture axée sur la compréhension et la gestion des risques. Il incombe aux utilisateurs du Cadre de continuer à exécuter la fonction **GOVERN** à mesure que les connaissances, les cultures et les besoins ou attentes des acteurs de l'IA évoluent au fil du temps.

Les pratiques liées à la gestion des risques liés à l'IA sont décrites dans le NIST AI RMF Playbook. Le tableau 1 présente les catégories et sous-catégories de la fonction **GOVERN**.

Tableau 1 : Catégories et sous-catégories de la fonction **GOUVERN**.

Catégories	Sous-catégories
GOUVERNEMENT 1 : Les politiques, processus, procédures et pratiques de l'organisation concernant la cartographie, la mesure et la gestion des risques liés à l'IA sont en place, transparents et mis en œuvre de manière efficace.	**GOUVERNANCE 1.1 :** Les exigences légales et réglementaires en matière d'IA sont comprises, gérées et documentées.
	GOUVERNEMENT 1.2 : Les caractéristiques d'une IA digne de confiance sont intégrées dans les politiques, les processus, les procédures et les pratiques de l'organisation.
	GOUVERNEMENT 1.3 : Des processus, des procédures et des pratiques sont en place pour déterminer le niveau nécessaire d'activités de gestion des risques en fonction de la tolérance au risque de l'organisation.
	GOUVERNEMENT 1.4 : Le processus de gestion des risques et ses résultats sont établis au moyen de politiques, de procédures et d'autres contrôles transparents fondés sur les priorités de l'organisation en matière de risques.

Suite à la page suivante

Tableau 1 : Catégories et sous-catégories de la fonction **GOUVERN**. (suite)

Catégories	Sous-catégories
	GOUVERNEMENT 1.5 : Le suivi permanent et l'examen périodique du processus de gestion des risques et de ses résultats sont planifiés et les rôles et responsabilités de l'organisation sont clairement définis, y compris la détermination de la fréquence de l'examen périodique. **GOUVERNEMENT 1.6 :** Des mécanismes sont en place pour inventorier les systèmes d'IA et sont dotés de ressources en fonction des priorités de l'organisation en matière de risques. **GOUVERNEMENT 1.7 :** Des processus et des procédures sont en place pour démanteler et supprimer progressivement les systèmes d'IA en toute sécurité et d'une manière qui n'augmente pas les risques et ne diminue pas la fiabilité de l'organisation.
GOUVERNEMENT 2 : Des structures de responsabilité sont en place afin que les équipes et les personnes appropriées soient habilitées, responsables et formées à la cartographie, à la mesure et à la gestion des risques liés à l'IA.	**GOUVERNEMENT 2.1 :** Les rôles, les responsabilités et les lignes de communication liés à la cartographie, à la mesure et à la gestion des risques liés à l'IA sont documentés et clairs pour les personnes et les équipes dans l'ensemble de l'organisation. **GOUVERNEMENT 2.2 :** Le personnel et les partenaires de l'organisation reçoivent une formation à la gestion des risques liés à l'IA qui leur permet de s'acquitter de leurs tâches et de leurs responsabilités conformément aux politiques, procédures et accords en la matière. **GOUVERNEMENT 2.3 :** La direction de l'organisation assume la responsabilité des décisions relatives aux risques associés au développement et au déploiement des systèmes d'IA.
GOUVERNEMENT 3 : Les processus de diversité, d'équité, d'inclusion et d'accessibilité du personnel sont prioritaires dans la cartographie, la mesure et la gestion des risques liés à l'IA tout au long du cycle de vie.	**GOUVERNEMENT 3.1 :** La prise de décision relative à la cartographie, à la mesure et à la gestion des risques liés à l'IA tout au long du cycle de vie est éclairée par une équipe diversifiée (diversité des caractéristiques démographiques, des disciplines, de l'expérience, de l'expertise et des antécédents, par exemple). **GOUVERNANCE 3.2 :** Des politiques et des procédures sont en place pour définir et différencier les rôles et les responsabilités en matière de configuration de l'IA par l'homme et de supervision des systèmes d'IA.
GOUVERNEMENT 4 : Les équipes organisationnelles s'engagent en faveur d'une culture	**GOUVERNE 4.1 :** Des politiques et des pratiques organisationnelles sont en place pour encourager la pensée critique et l'esprit de sécurité dans la conception, le développement, le déploiement et l'utilisation des systèmes d'IA afin de minimiser les impacts négatifs potentiels.

Suite à la page suivante

Tableau 1 : Catégories et sous-catégories de la fonction **GOUVERN**. (suite)

Catégories	Sous-catégories
qui prend en compte et communique les risques liés à l'IA.	**GOUVERNEMENT 4.2 :** Les équipes organisationnelles documentent les risques et les impacts potentiels de la technologie d'IA qu'elles conçoivent, développent, déploient, évaluent et utilisent, et elles communiquent plus largement sur ces impacts. **GOUVERNEMENT 4.3 :** Des pratiques organisationnelles sont en place pour permettre de tester l'IA, d'identifier les incidents et de partager les informations.
GOUVERNEMENT 5 : Des processus sont en place pour un engagement solide avec les acteurs de l'IA concernés.	**GOUVERNEMENT 5.1 :** Des politiques et des pratiques organisationnelles sont en place pour recueillir, examiner, hiérarchiser et intégrer les réactions des personnes extérieures à l'équipe qui a développé ou déployé le système d'IA en ce qui concerne les incidences individuelles et sociétales potentielles liées aux risques de l'IA. **GOUVERNEMENT 5.2 :** Des mécanismes sont mis en place pour permettre à l'équipe qui a développé ou déployé les systèmes d'IA d'intégrer régulièrement dans la conception et la mise en œuvre du système le retour d'information des acteurs concernés par l'IA.
GOUVERNEMENT 6 : Des politiques et des procédures sont mises en place pour gérer les risques et les avantages liés à l'IA découlant des éléments suivants les logiciels et les données de tiers et d'autres problèmes liés à la chaîne d'approvisionnement.	**GOUVERNANCE 6.1 :** Des politiques et des procédures sont en place pour gérer les risques d'IA associés aux entités tierces, y compris les risques d'atteinte à la propriété intellectuelle ou à d'autres droits d'un tiers. **GOUVERNANCE 6.2 :** Des procédures d'urgence sont en place pour gérer les défaillances ou les incidents dans les systèmes de données ou d'intelligence artificielle de tiers considérés présentant un risque élevé.

5.2 Carte

La fonction **MAP** établit le contexte permettant d'encadrer les risques liés à un système d'IA. Le cycle de vie de l'IA se compose de nombreuses activités interdépendantes impliquant un ensemble diversifié d'acteurs (voir figure 3). Dans la pratique, les acteurs de l'IA chargés d'une partie du processus n'ont souvent pas une visibilité ou un contrôle total sur les autres parties et leurs contextes associés. Les interdépendances entre ces activités et entre les acteurs de l'IA concernés peuvent rendre difficile l'anticipation fiable des impacts des systèmes d'IA. Par , les décisions prises au début de l'identification des buts et objectifs d'un système d'IA peuvent modifier son comportement et ses capacités, et la dynamique du cadre de déploiement (comme les utilisateurs finaux ou les personnes concernées) peut façonner les incidences des décisions relatives aux systèmes d'IA. Par conséquent, les meilleures intentions dans une dimension du cycle de vie de l'IA peuvent compromises par des interactions avec des décisions et des conditions dans d'autres activités ultérieures.

Cette complexité et les différents niveaux de visibilité peuvent introduire de l'incertitude dans les pratiques de gestion des risques. L'anticipation, l'évaluation et la prise en compte des sources potentielles de risque négatif peuvent atténuer cette incertitude et renforcer l'intégrité du processus décisionnel.

Les informations recueillies au cours de l'exécution de la fonction **MAP** permettent une prévention négative des risques et éclairent les décisions relatives à des processus tels que la gestion des modèles, ainsi qu'une décision initiale sur l'opportunité ou la nécessité d'une solution d'IA. Les résultats de la fonction **MAP** constituent la base des fonctions **MEASURE** et **MANAGE**. Sans connaissance du contexte et sans conscience des risques dans les contextes identifiés, la gestion des risques est difficile à mettre en œuvre. La fonction **MAP** vise à améliorer la capacité d'une organisation à identifier les risques et les facteurs contributifs plus généraux.

La mise en œuvre de cette fonction est renforcée par l'intégration des perspectives d'une équipe interne diversifiée et par l'engagement de personnes extérieures à l'équipe qui a développé ou déployé le système d'IA. L'implication des collaborateurs externes, des utilisateurs finaux, des communautés potentiellement touchées et d'autres personnes peut varier en fonction du niveau de risque d'un système d'IA particulier, de la composition de l'équipe interne et des politiques de l'organisation. La collecte de perspectives aussi larges peut aider les organisations à prévenir de manière proactive les risques négatifs et à développer des systèmes d'IA plus fiables :

- améliorer leur capacité à comprendre les contextes ;
- vérifier leurs hypothèses sur le contexte d'utilisation ;
- permettant de reconnaître les cas où les systèmes ne sont pas fonctionnels dans leur contexte ou en dehors de celui-ci
- identifier les utilisations positives et bénéfiques de leurs systèmes d'IA existants ;
- améliorer la compréhension des limites des processus d'IA et de ML ;
- identifier les contraintes dans les applications du monde réel qui peuvent avoir des conséquences négatives ;
- identifier les impacts négatifs connus et prévisibles liés à l'utilisation prévue des systèmes d'IA ; et
- anticiper les risques liés à l'utilisation des systèmes d'IA au-delà de l'usage prévu.

À l'issue de la fonction **MAP**, les utilisateurs du cadre devraient disposer de connaissances contextuelles suffisantes sur les impacts des systèmes d'IA pour prendre une décision initiale d'acceptation ou de refus concernant la conception, le développement ou le déploiement d'un système d'IA. Si décision est prise, les organisations doivent utiliser les fonctions **MESURER** et **GÉRER**, ainsi que les politiques et procédures mises en place dans le cadre de la fonction **GOUVERNER**, afin de contribuer aux efforts de gestion des risques liés à l'IA. Les utilisateurs du cadre sont tenus de continuer à appliquer la fonction **MAP** aux systèmes d'IA à mesure que le contexte, les capacités, les risques, les avantages et les incidences potentielles évoluent dans le temps.

Les pratiques liées à la cartographie des risques liés à l'IA sont décrites dans le NIST AI RMF Playbook. Le tableau 2 énumère les catégories et sous-catégories de la fonction **MAP.**

Tableau 2 : Catégories et sous-catégories de la fonction **MAP.**

Catégories	Sous-catégories
MAP 1 : Le contexte est établi et compris.	**MAP 1.1 :** Les objectifs visés, les utilisations potentiellement bénéfiques, les lois, normes et attentes spécifiques au contexte, ainsi que les contextes potentiels dans lesquels le système d'IA sera déployé sont compris et documentés. Les considérations comprennent : l'ensemble ou les types spécifiques d'utilisateurs ainsi que leurs attentes ; les incidences positives et négatives potentielles des utilisations du système sur les individus, les communautés, les organisations, la société et la planète ; les hypothèses et les limites connexes concernant les objectifs, les utilisations et les risques du système d'IA tout au long du cycle de développement ou de vie de l'IA du produit ; la VET et les mesures du système connexes.
	MAP 1.2 : Les acteurs, les compétences, les aptitudes et les capacités interdisciplinaires en matière d'IA pour établir le contexte reflètent la diversité démographique et une vaste expertise en matière de domaines et d'expérience des utilisateurs, et leur par- ticipation est documentée. Les possibilités de collaboration interdisciplinaire sont classées par ordre de priorité.
	MAP 1.3 : La mission de l'organisation et les objectifs pertinents en matière de technologie de l'IA sont compris et documentés.
	MAP 1.4 : La valeur commerciale ou le contexte de l'utilisation commerciale a été clairement défini ou - dans le cas de l'évaluation de systèmes d'IA existants - réévalué.
	MAP 1.5 : Les seuils de tolérance au risque de l'organisation sont déterminés et documentés.
	MAP 1.6 : Les exigences du système (par exemple, "le système doit respecter la vie privée de ses utilisateurs") sont demandées aux acteurs concernés de l'IA et comprises par eux. Les décisions en matière de conception tiennent compte des implications sociotechniques pour faire face aux risques liés à l'IA.
MAP 2 : La catégorisation du système d'IA est effectuée.	**MAP 2.1 :** Les tâches spécifiques et les méthodes utilisées pour mettre en œuvre tâches que le système d'IA soutiendra sont définies (par exemple, les classificateurs, les modèles génératifs, les recommandeurs).
	PAM 2.2 : Les informations relatives aux limites des connaissances du système d'IA et à la manière dont les résultats du système peuvent être utilisés et supervisés par des êtres humains sont documentées. La documentation fournit suffisamment d'informations pour aider les acteurs concernés de l'IA à prendre des décisions et des mesures ultérieures.

Suite à la page suivante

Tableau 2 : Catégories et sous-catégories de la fonction **MAP.** (suite)

Catégories	Sous-catégories
	MAP 2.3 : L'intégrité scientifique et les considérations relatives à la VET sont identifiées et documentées, y compris celles liées à la conception expérimentale, à la collecte et à la sélection des données (par exemple, disponibilité, représentativité, adéquation), à la fiabilité du système et à la validation de la construction.
MAP 3 : IA les capacités, l'utilisation ciblée, les objectifs, ainsi que les avantages et les coûts escomptés par rapport à des critères de référence appropriés sont compris.	**PAM 3.1 :** Les avantages potentiels des fonctionnalités et des performances prévues du système d'IA sont examinés et documentés. **MAP 3.2 :** Les coûts potentiels, y compris les coûts non monétaires, qui résultent d'erreurs d'IA prévues ou réalisées ou de la fonctionnalité et de la fiabilité du système - en rapport avec la tolérance au risque de l'organisation - sont examinés et documentés. **MAP 3.3 :** Le champ d'application ciblé est spécifié et documenté sur la base des capacités du système, du contexte établi et de la catégorisation des systèmes d'IA. **PAM 3.4 :** Les processus relatifs à la compétence des opérateurs et des praticiens en ce qui concerne les performances et la fiabilité des systèmes d'IA - ainsi que les normes techniques et les certifications correspondantes - sont définis, évalués et documentés. **MAP 3.5 :** Les processus de supervision humaine sont définis, évalués et documentés conformément aux politiques organisationnelles de la fonction **GOUVERN.**
MAP 4 : Les risques et les avantages sont cartographiés pour tous les composants du système d'IA, y compris les logiciels et les données de tiers.	**MAP 4.1 :** Des approches permettant de cartographier la technologie de l'IA et les risques juridiques liés à ses composantes - y compris l'utilisation de données ou de logiciels de tiers - sont en , suivies et documentées, de même que les risques d'atteinte à la propriété intellectuelle ou à d'autres droits d'un . **PAM 4.2 :** Les contrôles internes des risques pour les composantes du système d'IA, y compris les technologies d'IA de tiers, sont recensés et documentés.
MAP 5 : Les impacts sur les individus, les groupes, les communautés, les organisations et la société sont caractérisés.	**MAP 5.1 :** La probabilité et l'ampleur de chaque impact identifié (à la fois potentiellement bénéfique et nuisible) sur la base de l'utilisation prévue, des utilisations antérieures de systèmes d'IA dans des contextes similaires, des incidents publics, du retour d'information de la part des personnes extérieures à l'équipe qui a développé ou déployé le système d'IA, ou d'autres données, sont identifiées et documentées.

Suite à la page suivante

Tableau 2 : Catégories et sous-catégories de la fonction **MAP.** (suite)

Catégories	Sous-catégories
	PAM 5.2 : Des pratiques et du personnel sont en place et documentés pour soutenir un engagement régulier avec les acteurs concernés de l'IA et intégrer le retour d'information sur les impacts positifs, négatifs et imprévus.

5.3 Mesure

La fonction **MESURE** utilise des outils, des tech- niques et des méthodologies quantitatives, qualitatives ou mixtes pour analyser, évaluer, étalonner et surveiller les risques liés à l'IA et les incidences connexes. Elle utilise les connaissances relatives aux risques liés à l'IA identifiés dans le cadre de la fonction **MAP** et informe la fonction **MANAGE**. Les systèmes d'IA devraient être testés avant leur déploiement et régulièrement en cours d'exploitation. Les mesures des risques liés à l'IA comprennent la documentation des aspects de la fonctionnalité et de la fiabilité des systèmes.

La mesure des risques liés à l'IA comprend le suivi des paramètres relatifs aux caractéristiques de confiance, à l'impact social et aux configurations homme-AI. Les processus élaborés ou adoptés dans le cadre de la fonction **MESURE** devraient comprendre des méthodes rigoureuses de test des logiciels et d'évaluation des performances, avec des mesures d'incertitude associées, des comparaisons avec des critères de performance, ainsi que des rapports et une documentation formalisés des résultats. Des processus d'examen indépendant peuvent améliorer l'efficacité des essais et atténuer les préjugés internes et les conflits d'intérêts potentiels.

Lorsque des compromis entre les caractéristiques de confiance se présentent, les mesures fournissent une base traçable pour informer les décisions de gestion. Les options peuvent inclure le recalibrage, l'atténuation de l'impact ou le retrait du système de la conception, du développement, de la production ou de l'utilisation, ainsi 'une série de contrôles compensatoires, de détection, de dissuasion, de directive et de récupération.

Après avoir rempli la fonction **MESURE**, des processus de test, d'évaluation, de vérification et de validation (TEVV) objectifs, reproductibles ou évolutifs, comprenant des métriques, des méthodes et des méthodologies, sont mis en place, suivis et documentés. Les métriques et les méthodologies de mesure devraient respecter les normes scientifiques, juridiques et éthiques et être mises en œuvre dans le cadre d'un processus ouvert et trans-parental. Il peut s'avérer nécessaire de développer de nouveaux types de mesures, qualitatives et quantitatives. Il convient d'examiner dans quelle mesure chaque type de mesure fournit des informations uniques et significatives pour l'évaluation des risques liés à l'IA. Les utilisateurs du cadre amélioreront leur capacité à évaluer de manière exhaustive la fiabilité des systèmes, à identifier et à suivre les risques existants et émergents, et à vérifier l'efficacité des mesures. Les résultats des mesures seront utilisés dans le cadre de la fonction de **gestion (MANAGE)** pour faciliter la surveillance des risques et les efforts de réaction. Il incombe aux utilisateurs du cadre de continuer à appliquer la fonction **MESURE** aux systèmes d'IA à mesure que les connaissances, les méthodologies, les risques et les incidences évoluent dans le temps.

Les pratiques liées à la mesure des risques liés à l'IA sont décrites dans le NIST AI RMF Playbook. Le tableau 3 énumère les catégories et sous-catégories de la fonction **MEASURE.**

Tableau 3 : Catégories et sous-catégories de la fonction **MESURE.**

Catégories	Sous-catégories
MESURE 1 : Des méthodes et des mesures appropriées sont identifiées et appliquées.	**MESURE 1.1 :** Les approches et les paramètres de mesure des risques liés à l'IA énumérés dans le cadre de la fonction **MAP** sont sélectionnés en vue de leur mise en œuvre, en commençant par les risques liés à l'IA les plus importants. Les risques ou les caractéristiques de fiabilité qui ne seront pas - ou ne peuvent pas être - mesurés sont correctement documentés.
	MESURE 1.2 : L'adéquation des paramètres de l'IA et l'efficacité des contrôles existants sont régulièrement évaluées et mises à jour, y compris les rapports sur les erreurs et les incidences potentielles sur les communautés touchées.
	MESURE 1.3 : Des experts internes qui n'ont pas été des développeurs de première ligne pour le système et/ou des évaluateurs indépendants participent à des évaluations et à des mises à jour régulières. Les experts du domaine, les utilisateurs, les acteurs de l'IA extérieurs à l'équipe qui a développé ou déployé le système d'IA et les communautés concernées sont consultés à l'appui des évaluations, en fonction de la tolérance au risque de l'organisation.
MESURE 2 : AI sont évalués en fonction de leurs caractéristiques de fiabilité.	**MESURE 2.1 :** Les jeux de tests, les mesures et les détails des outils utilisés pendant la VET sont documentés.
	MESURE 2.2 : Les évaluations impliquant des sujets humains répondent aux exigences applicables (y compris la protection des sujets humains) et sont représentatives de la population concernée.
	MESURE 2.3 : Les performances du système d'IA ou les critères d'assurance sont mesurés qualitativement ou quantitativement et démontrés dans des conditions similaires au(x) contexte(s) de déploiement. Les mesures sont documentées.
	MESURE 2.4 : fonctionnalité et le comportement du système d'IA et de ses composants - tels qu'identifiés dans la fonction **MAP** - sont contrôlés en production.
	MESURE 2.5 : La validité et la fiabilité du système d'IA à déployer sont démontrées. Les limites de la généralisation au-delà des conditions dans lesquelles la technologie a été développée sont documentées.

Suite à la page suivante

Tableau 3 : Catégories et sous-catégories de la fonction **MESURE**. (suite)

Catégories	Sous-catégories
	MESURE 2.6 : Le système d'IA est régulièrement évalué en ce qui concerne les risques pour la sécurité, tels qu'ils sont identifiés dans la fonction **MAP**. Il est démontré que le système d'IA à déployer est sûr, que son risque négatif résiduel ne dépasse pas la tolérance au risque et qu'il peut tomber en panne en toute sécurité, en particulier s'il est amené à fonctionner au-delà des limites de ses connaissances. Les mesures de sécurité reflètent la fiabilité et la robustesse du système, la surveillance en temps réel et les délais de réaction en cas de défaillance du système d'IA.
	MESURE 2.7 : La sécurité et la résilience du système d'IA - telles qu'elles sont définies dans la fonction **MAP** - sont évaluées et documentées.
	MESURE 2.8 : Les risques liés à la transparence et à la capacité de rendre compte - tels qu'ils sont identifiés dans le cadre de la fonction de **MAP** - sont examinés et documentés.
	MESURE 2.9 : Le modèle d'IA est expliqué, validé et documenté, et les résultats du système d'IA sont interprétés dans leur contexte - tel qu'il est défini dans la fonction **MAP** - afin de permettre une utilisation et une gouvernance responsables.
	MESURE 2.10 : Le risque d'atteinte à la vie privée lié au système d'IA - tel qu'il est identifié dans la fonction **MAP** - est examiné et documenté.
	MESURE 2.11 : Équité et partialité - telles qu'identifiées dans la **MAP** sont évalués et les résultats sont documentés.
	MESURE 2.12 : L'impact sur l'environnement et la durabilité des activités de formation et de gestion des modèles d'IA - telles qu'identifiées dans la fonction **MAP** - sont évalués et documentés.
	MESURE 2.13 : L'efficacité des et processus TEVV employés dans la fonction **MESURE** est évaluée et documentée.
MESURE 3 : Des mécanismes de suivi des risques d'IA identifiés au fil du temps sont en place.	**MESURE 3.1 :** Des approches, du personnel et de la documentation sont en place pour identifier et suivre régulièrement les risques existants, imprévus et émergents en matière d'IA, sur la base de facteurs tels que les performances prévues et réelles dans des contextes de déploiement.
	MESURE 3.2 : Des approches de suivi des risques sont envisagées dans les contextes où les risques liés à l'IA sont difficiles à évaluer à l'aide des techniques de mesure actuellement disponibles ou lorsque les paramètres ne sont pas encore disponibles.

Suite à la page suivante

Tableau 3 : Catégories et sous-catégories de la fonction **MESURE.** (suite)

Catégories	Sous-catégories
	MESURE 3.3 : Des processus de retour d'information permettant aux utilisateurs finaux et aux communautés concernées de signaler les problèmes et de faire appel des résultats du système sont mis en place et intégrés dans les mesures d'évaluation du système d'IA.
MESURE 4 : Le retour d'information sur l'efficacité de la mesure est recueilli et évalué.	**MESURE 4.1 :** Les méthodes de mesure permettant d'identifier les risques liés à l'IA sont liées au(x) contexte(s) de déploiement et s'appuient sur la consultation d'experts du domaine et d'autres utilisateurs finaux. Les approches sont documentées. **MESURE 4.2 :** Les résultats des mesures concernant la fiabilité des systèmes d'IA dans le(s) contexte(s) de déploiement et tout au long du cycle de vie de l'IA s'appuient sur les contributions des experts du domaine et des acteurs de l'IA concernés afin de valider si le système fonctionne systématiquement comme prévu. Les résultats sont documentés. **MESURE 4.3 :** Des améliorations ou des dé- clins mesurables des performances, fondés sur des consultations avec les acteurs concernés de l'IA, y compris les communautés affectées, et sur des données de terrain concernant les risques et les caractéristiques de fiabilité propres au contexte, sont identifiés et documentés.

5.4 Gérer

La fonction **GÉRER** consiste à allouer des ressources aux risques cartographiés et mesurés sur une base régulière et selon les modalités définies par la fonction **GOUVERNER**. Le traitement des risques comprend des plans visant à répondre aux incidents ou aux événements, à s'en remettre et à communiquer à leur sujet.

Les informations contextuelles obtenues grâce à la consultation d'experts et à la contribution des acteurs concernés par l'IA - établies dans **GOVERN** et réalisées dans **MAP** - sont utilisées dans cette fonction pour réduire la probabilité de défaillances du système et d'incidences négatives. Les pratiques de documentation systématique établies dans **GOVERN** et utilisées dans **MAP** et **MEASURE** soutiennent les efforts de gestion des risques liés à l'IA et augmentent la transparence et la responsabilité. Des processus d'évaluation des risques émergents sont en place, ainsi que des mécanismes d'amélioration continue.

Une fois la fonction de **gestion** achevée, des plans de hiérarchisation des risques, de suivi régulier et d'amélioration seront mis en . Les utilisateurs du cadre une meilleure capacité à gérer les risques des systèmes d'IA déployés et à allouer des ressources de gestion des risques en fonction des risques évalués et hiérarchisés. Il incombe aux utilisateurs du cadre de continuer à appliquer la fonction de **gestion** aux systèmes d'IA déployés, à mesure que les méthodes, les contextes, les risques et les besoins ou attentes des acteurs concernés par l'IA évoluent au fil du temps.

Les pratiques liées à la gestion des risques liés à l'IA sont décrites dans le NIST AI RMF Playbook. Le tableau 4 énumère les catégories et sous-catégories de la fonction **GÉRER.**

Tableau 4 : Catégories et sous-catégories de la fonction **GÉRER.**

Catégories	Sous-catégories
GÉRER 1 : AI sur la base d'évaluations et d'autres résultats analytiques provenant de l'Agence européenne pour la sécurité et la santé au travail. **CARTE** et **MESURE** sont hiérarchisées, font l'objet d'une réponse et sont gérées.	**GÉRER 1.1 :** Il est déterminé si le système d'IA atteint les buts visés et les objectifs déclarés et s'il convient de poursuivre son développement ou son déploiement. **GÉRER 1.2 :** Le traitement des risques documentés liés à l'IA est hiérarchisé en fonction de l'impact, de la probabilité et des ressources ou méthodes disponibles. **GÉRER 1.3 :** Les réponses aux risques liés à l'IA jugés hautement prioritaires, tels qu'identifiés par la fonction **MAP,** sont élaborées, planifiées et documentées. Les options de réponse aux risques peuvent comprendre l'atténuation, le transfert, l'évitement ou l'acceptation. **GÉRER 1.4 :** Les risques résiduels négatifs (définis comme la somme de tous les risques non atténués) pour les acquéreurs en aval des systèmes d'IA et les utilisateurs finaux sont documentés.
GÉRER 2 : Les stratégies visant à maximiser les bénéfices de l'IA et à minimiser les impacts négatifs sont planifiées, préparées, mises en œuvre, documentées et informées par les acteurs concernés de l'IA.	**GÉRER 2.1 :** Les ressources nécessaires à la gestion des risques liés à l'IA sont prises en compte - de même que les systèmes, approches ou méthodes de rechange viables non liés à l'IA - afin de réduire l'ampleur ou la probabilité des incidences potentielles. **GÉRER 2.2 :** Des mécanismes sont en place et appliqués pour maintenir la valeur des systèmes d'IA déployés. **GÉRER 2.3 :** Des procédures sont suivies pour répondre à un risque précédemment inconnu et s'en remettre lorsqu'il est identifié. **GÉRER 2.4 :** Des mécanismes sont en place et appliqués, et les responsabilités sont attribuées et comprises, pour remplacer, désengager ou désactiver les systèmes d'IA dont les performances ou les résultats sont incompatibles avec l'utilisation prévue.
GÉRER 3 : IA les risques et les avantages provenant d'entités tierces sont gérés.	**GÉRER 3.1 :** Les risques et les avantages liés à l'IA provenant de ressources tierces font l'objet d'un suivi régulier, et des contrôles des risques sont appliqués et documentés. **GÉRER 3.2 :** Les modèles pré-entraînés utilisés pour le développement sont contrôlés dans le cadre de la surveillance et de la maintenance régulières du système d'IA.

Suite à la page suivante

Tableau 4 : Catégories et sous-catégories de la fonction **GÉRER.** (suite)

Catégories	Sous-catégories
GÉRER 4 : Risques Les traitements, y compris les plans d'intervention et de récupération, ainsi que les plans de communication pour les risques d'IA identifiés et mesurés sont documentés et contrôlés régulièrement.	**GÉRER 4.1 :** Des plans de surveillance du système d'IA après le déploiement sont mis en œuvre, y compris des mécanismes de collecte et d'évaluation des données fournies par les utilisateurs et d'autres acteurs concernés par l'IA, des mécanismes d'appel et d'annulation, de déclassement, de réponse aux incidents, de récupération et de gestion des changements. **GÉRER 4.2 :** Des activités mesurables d'amélioration continue sont intégrées dans les mises à jour du système d'IA et comprennent un engagement régulier avec les parties intéressées, y compris les acteurs pertinents de l'IA. **GÉRER 4.3 :** Les incidents et les erreurs sont communiqués aux acteurs concernés de l'IA, y compris aux communautés affectées. Les processus de suivi, de réaction et de récupération des incidents et des erreurs sont suivis et documentés.

6. Profils AI RMF

Les profils de cas d'utilisation de l'AI RMF sont des mises en œuvre des fonctions, catégories et sous-catégories de l'AI RMF dans un cadre ou une application spécifique, en fonction des exigences, de la tolérance au risque et des ressources de l'utilisateur du cadre : par exemple, un *profil d'embauche de l'*AI RMF ou un *profil de logement équitable de l'*AI RMF. Les profils peuvent illustrer et donner un aperçu de la manière dont les risques peuvent être gérés à différents stades du cycle de vie de l'IA ou dans des secteurs, des technologies ou des applications finales spécifiques. Les profils des CGR de l'IA aident les organisations à décider de la meilleure façon de gérer les risques liés à l'IA en fonction de leurs objectifs, des exigences légales/réglementaires et des meilleures pratiques, et des priorités en matière de gestion des risques.

Les profils temporels du cadre de référence de l'IA sont des descriptions de l'état actuel ou de l'état cible souhaité d'activités spécifiques de gestion des risques liés à l'IA dans un secteur, une industrie, une organisation ou un contexte d'application donné. Un profil actuel du cadre de gestion des risques liés à l'IA indique comment l'IA est actuellement gérée et quels sont les risques associés en termes de résultats actuels. Un profil cible indique les résultats nécessaires pour atteindre les objectifs souhaités ou cibles en matière de gestion des risques liés à l'IA.

La comparaison entre le profil actuel et le profil cible révèle probablement des lacunes à combler pour atteindre les objectifs de gestion des risques liés à l'IA. Des plans d'action peuvent être élaborés pour combler ces lacunes et atteindre les résultats dans une catégorie ou sous-catégorie donnée. L'ordre de priorité de l'atténuation des lacunes dépend des besoins de l'utilisateur et de ses processus de gestion des risques. Cette approche basée sur les risques permet également aux utilisateurs du cadre de comparer leurs approches avec d'autres approches et d'évaluer les ressources nécessaires (par exemple, le personnel, le financement) pour atteindre les objectifs de gestion des risques liés à l'IA d'une manière rentable et en établissant des priorités.

Les *profils intersectoriels* du RMF IA couvrent les risques des modèles ou des applications qui peuvent être utilisés dans différents cas d'utilisation ou secteurs. Les profils intersectoriels peuvent également couvrir la manière de gouverner, de cartographier, de mesurer et de gérer les risques pour les activités ou les processus commerciaux communs à plusieurs secteurs, tels que l'utilisation de grands modèles linguistiques, les services basés sur le cloud ou l'acquisition.

Le présent cadre ne prescrit pas de modèles de profil, ce qui permet une certaine souplesse dans la mise en œuvre.

Annexe A :
Description des tâches des acteurs de l'IA dans les figures 2 et 3

Les tâches de **conception de l'IA** sont exécutées au cours des phases Contexte de l'application et Données et entrées du cycle de vie de l'IA (figure 2). Les acteurs de la conception de l'IA créent le concept et les objectifs des systèmes d'IA et sont responsables des tâches de planification, de conception, de collecte et de traitement des données du système d'IA afin que ce dernier soit légal et adapté à son objectif. Les tâches comprennent l'ar- ticulation et la documentation du concept et des objectifs du système, des hypothèses sous-jacentes, du contexte et des exigences ; la collecte et le nettoyage des données ; et la documentation des métadonnées et des caractéristiques de l'ensemble de données. Les acteurs de l'IA dans cette catégorie comprennent les scientifiques des données, les experts en do- main, les analystes socioculturels, les experts dans le domaine de la diversité, de l'équité, de l'inclusion et de l'accessibilité, les membres des communautés touchées, les experts en facteurs humains (par exemple, conception UX/UI), les experts en gouvernance, les ingénieurs des données, les fournisseurs de données, les financeurs du système, les gestionnaires de produits, les entités tierces, les évaluateurs et la gouvernance juridique et de la protection de la vie privée.

Les tâches de **développement de l'IA** sont exécutées pendant la phase du cycle de vie consacrée au modèle d'IA, comme le montre la figure suivante
2. Les acteurs du développement de l'IA fournissent l'infrastructure initiale des systèmes d'IA et sont responsables des tâches d'élaboration et d'interprétation des modèles, qui impliquent la création, la sélection, la cali- bration, l'entraînement et/ou le test de modèles ou d'algorithmes. Les acteurs de l'IA de cette catégorie comprennent des experts en apprentissage automatique, des scientifiques des données, des développeurs, des entités tierces, des experts en gouvernance juridique et en protection de la vie privée, ainsi que des experts des facteurs socioculturels et contextuels associés au contexte de déploiement.

Les tâches de **déploiement de l'IA** sont exécutées au cours de la phase des tâches et des résultats du cycle de vie (figure 2). Les acteurs du déploiement de l'IA sont responsables des décisions contextuelles relatives à l'utilisation du système d'IA afin d'assurer le déploiement du système en production. Les tâches connexes comprennent le pilotage du système, la vérification de la compatibilité avec les systèmes existants, la garantie de la conformité réglementaire, la gestion des changements organisationnels et l'évaluation de l'expérience de l'utilisateur. Les acteurs de l'IA dans cette catégorie comprennent les intégrateurs de systèmes, les développeurs de logiciels, les utilisateurs finaux, les opérateurs et les praticiens, les évaluateurs et les experts du domaine ayant une expertise en matière de facteurs humains, d'analyse socioculturelle et de gouvernance.

Les tâches d'**exploitation et de surveillance** sont exécutées dans la phase Contexte de l'application/exploitation et surveillance du cycle de vie de la figure 2. Ces tâches sont exécutées par des acteurs de l'IA qui sont chargés d'exploiter le système d'IA et de collaborer avec d'autres pour évaluer régulièrement les résultats et les incidences du système. Les acteurs de l'IA appartenant à cette catégorie comprennent les opérateurs du système, les experts du domaine, les concepteurs de l'IA, les utilisateurs qui interprètent ou intègrent les résultats des systèmes d'IA, les développeurs de produits, les évaluateurs et les auditeurs, les experts en conformité, la direction de l'organisation et les membres de la communauté des chercheurs.

Les tâches de **test, d'évaluation, de vérification et de validation (TEVV)** sont exécutées tout au long du cycle de vie de l'IA. Elles sont exécutées par des acteurs de l'IA qui examinent le système d'IA ou ses composants, ou qui détectent et corrigent les problèmes. Idéalement, les acteurs de l'IA qui effectuent la vérification

Les tâches d'évaluation et de validation sont distinctes des d'essai et d'évaluation. Les tâches peuvent être intégrées dans une phase dès la conception, où les tests sont planifiés conformément aux exigences de conception.

- Les tâches de la TEVV relatives à la conception, à la planification et aux données peuvent être centrées sur la validation interne et externe des hypothèses relatives à la conception du système, à la collecte des données et aux mesures par rapport au contexte de déploiement ou d'application prévu.
- Les tâches de TEVV pour le développement (c'est-à-dire la construction du modèle) comprennent la validation et l'évaluation du modèle.
- Les tâches de TEVV pour le déploiement comprennent la validation et l'intégration du système dans la production, avec des tests et un recalibrage pour l'intégration des systèmes et des processus, l'expérience de l'utilisateur et la conformité avec les spécifications légales, réglementaires et éthiques existantes.
- Les tâches de VTEP pour les opérations impliquent un contrôle continu pour les mises à jour périodiques, les tests et le recalibrage des modèles par les experts en la matière, le suivi des incidents ou des erreurs signalés et leur gestion, la détection des propriétés émergentes et des incidences connexes, ainsi que les processus de réparation et de réaction.

Les tâches et activités liées **aux facteurs humains** se retrouvent dans toutes les dimensions du cycle de vie de l'IA. Elles comprennent des pratiques et des méthodologies de conception centrées sur l'homme, la promotion de la participation active des utilisateurs finaux et d'autres parties intéressées et acteurs pertinents de l'IA, l'intégration de normes et de valeurs spécifiques au contexte dans la conception du système, l'évaluation et l'adaptation des expériences des utilisateurs finaux, et l'intégration générale des humains et de la dynamique humaine dans toutes les phases cycle de vie de l'IA. Les professionnels des facteurs humains apportent des compétences et des perspectives multidisciplinaires pour comprendre le contexte d'utilisation, informer la diversité interdisciplinaire et démographique, s'engager dans des processus consultatifs, concevoir et évaluer l'expérience de l'utilisateur, effectuer des évaluations et des tests centrés sur l'homme et informer les évaluations d'impact.

Les tâches d'**expert de domaine** impliquent la contribution de praticiens ou d'universitaires multidisciplinaires qui fournissent des connaissances ou une expertise dans - et sur - un secteur industriel, un secteur économique, un con- texte ou un domaine d'application dans lequel un système d'IA est utilisé. Les acteurs de l'IA qui sont des experts du domaine peuvent fournir des orientations essentielles pour la conception et le développement des systèmes d'IA, et des résultats inter- prétaires à l'appui des travaux réalisés par les équipes de TEVV et d'évaluation de l'impact de l'IA.

Les tâches d'**évaluation de l'impact de l'IA** comprennent l'évaluation des exigences en matière de responsabilité des systèmes d'IA, la lutte contre les préjugés nuisibles, l'examen de l'impact des systèmes d'IA, la sécurité des produits, la responsabilité et la sécurité, entre autres. Les acteurs de l'IA, tels que les évaluateurs d'impact et les évaluateurs, fournissent une expertise technique, humaine, socioculturelle et juridique.

Les tâches d'**acquisition** sont menées par des acteurs de l'IA disposant d'une autorité de gestion financière, juridique ou politique pour l'acquisition de modèles, de produits ou de services d'IA auprès d'un développeur, d'un vendeur ou d'un contractant tiers.

Les tâches de **gouvernance et de supervision** sont assumées par des acteurs de l'IA ayant une autorité et une responsabilité de gestion, fiduciaire et juridique pour l'organisation dans laquelle un système

d'IA est utilisé.

signé, développé et/ou déployé. Les principaux acteurs de l'IA responsables de la gouvernance de l'IA sont la direction de l'organisation, les cadres supérieurs et le conseil d'administration. Ces acteurs sont les parties concernées par l'impact et la durabilité de l'organisation dans son ensemble.

Acteurs supplémentaires de l'IA

Les entités tierces comprennent les fournisseurs, les développeurs, les vendeurs et les évaluateurs de données, d'al- gorithmes, de modèles et/ou de systèmes et de services connexes pour une autre organisation ou pour les clients de l'orga- nisation. Les entités tierces sont responsables des tâches de conception et de développement de l'IA, en tout ou en partie. Par définition, elles sont extérieures à l'équipe de conception, de développement ou de déploiement de l'organisation qui acquiert ses technologies ou ses services. Les technologies acquises auprès d'entités tierces peuvent être complexes ou opaques, et la tolérance au risque peut ne pas correspondre à celle de l'organisation qui les déploie ou les exploite.

Les utilisateurs finaux d'un système d'IA sont les personnes ou les groupes qui utilisent le système à des fins spécifiques. Ces personnes ou groupes interagissent avec un système d'IA dans un contexte spécifique. Les compétences des utilisateurs finaux peuvent aller de celles d'experts en IA à celles d'utilisateurs novices en matière de technologie.

Les personnes/communautés concernées englobent tous les individus, groupes, communautés ou organisations directement ou indirectement concernés par les systèmes d'IA ou les décisions basées sur les résultats des systèmes d'IA. Ces personnes n'interagissent pas nécessairement avec le système ou l'application déployé.

D'autres acteurs de l'IA peuvent fournir des normes ou des orientations formelles ou quasi-formelles pour spécifier et gérer les risques liés à l'IA. Il peut s'agir d'**associations commerciales, d'organismes de normalisation, de groupes de pression, de chercheurs, de groupes environnementaux et d'organisations de la société civile**.

Le grand public est le plus susceptible de subir directement les effets positifs et négatifs des technologies de l'IA. Il peut motiver les mesures prises par les acteurs de l'IA. Ce groupe peut comprendre des individus, des communautés et des consommateurs associés au contexte dans lequel un système d'IA est développé ou déployé.

Annexe B :
En quoi les risques liés à l'IA diffèrent-ils des risques logiciels traditionnels ?

Comme pour les logiciels traditionnels, les risques liés aux technologies basées sur l'IA peuvent dépasser le cadre d'une entreprise, s'étendre à d'autres organisations et avoir des répercussions sur la société. Les systèmes d'IA comportent également un ensemble de risques qui ne sont pas pris en compte de manière exhaustive par les cadres et approches actuels en matière de risques. Certaines caractéristiques des systèmes d'IA qui présentent des risques peuvent également être bénéfiques. Par exemple, les modèles préformés et l'apprentissage par transfert peuvent faire progresser la recherche et accroître la précision et la résilience par rapport à d'autres modèles et approches. L'identification des facteurs contextuels dans la fonction **MAP** aidera les acteurs de l'IA à déterminer le niveau de risque et les efforts de gestion potentiels.

Par rapport aux logiciels traditionnels, les risques spécifiques à l'IA qui sont nouveaux ou accrus sont les suivants :

- Les données utilisées pour construire un système d'IA peuvent ne pas être une représentation fidèle ou appropriée du contexte ou de l'utilisation prévue du système d'IA, et la vérité de terrain peut ne pas exister ou ne pas être disponible. En outre, des biais nuisibles et d'autres problèmes de qualité des données peuvent affecter la fiabilité du système d'IA, ce qui pourrait avoir des conséquences négatives.
- La dépendance des systèmes d'IA à l'égard des données pour les tâches de formation, combinée à l'augmentation du volume et de la complexité généralement associés à ces données.
- Des changements intentionnels ou non intentionnels au cours de la formation peuvent altérer fondamentalement les performances des systèmes d'IA.
- Les ensembles de données utilisés pour former les systèmes d'intelligence artificielle peuvent être détachés de leur contexte original ou devenir périmés par rapport au contexte de déploiement.
- L'échelle et la complexité des systèmes d'IA (de nombreux systèmes contiennent des milliards, voire des trillions de points de décision) hébergés dans des applications logicielles plus traditionnelles.
- L'utilisation de modèles pré-entraînés, qui peut faire progresser la recherche et améliorer les performances, peut également augmenter les niveaux d'incertitude statistique et poser des problèmes de gestion des biais, de validité scientifique et de reproductibilité.
- Plus grande difficulté à prédire les modes de défaillance pour les propriétés émergentes des modèles pré-entraînés à grande échelle.
- Risque pour la vie privée en raison de la capacité accrue d'agrégation de données des systèmes d'IA.
- Les systèmes d'IA peuvent nécessiter une maintenance plus fréquente et des déclencheurs de maintenance corrective en raison de la dérive des données, des modèles ou des concepts.
- Opacité accrue et inquiétudes quant à la reproductibilité.
- Des normes de test des logiciels sous-développées et l'incapacité de documenter les pratiques basées sur l'IA selon les normes attendues des logiciels conçus de manière traditionnelle, sauf dans les cas les plus simples.
- Difficulté d'effectuer des tests réguliers de logiciels basés sur l'IA, ou de déterminer ce qu'il faut tester, car les systèmes d'IA ne sont pas soumis aux mêmes contrôles que le développement de codes traditionnels.

- Les coûts de calcul pour le développement de systèmes d'IA et leur impact sur l'environnement et la planète.
- L'incapacité à prédire ou à détecter les effets secondaires des systèmes basés sur l'IA au-delà des mesures statistiques.

Les considérations et les approches relatives à la gestion des risques en matière de protection de la vie privée et de cybersécurité s'appliquent à la conception, au développement, au déploiement, à l'évaluation et à l'utilisation des systèmes d'IA. Les risques en matière de protection de la vie privée et de cybersécurité sont également pris en compte dans le cadre de considérations plus larges gestion des risques de l'entreprise, qui peuvent intégrer les risques liés à l'IA. Dans le cadre des efforts déployés pour prendre en compte les caractéristiques de fiabilité de l'IA telles que "sécurisée et résiliente" et "respectueuse de la vie privée", les organisations peuvent envisager de s'appuyer sur les normes et orientations disponibles qui fournissent des orientations générales aux organisations pour réduire les risques en matière de sécurité et de respect de la vie privée, telles , sans s'y limiter, le cadre de cybersécurité du NIST, le cadre de respect de la vie privée du NIST, le cadre de gestion des risques du NIST et le cadre de développement de logiciels sécurisés. Ces cadres ont certaines caractéristiques en commun avec le cadre de gestion des risques de l'IA. Comme la plupart des approches de gestion des risques, ils sont axés sur les résultats plutôt que sur la prescription et sont souvent structurés autour d'un ensemble de fonctions, de catégories et de sous-catégories. Bien qu'il existe des différences significatives entre ces cadres en fonction du domaine traité - et parce que la gestion des risques liés à l'IA exige de prendre en compte de nombreux autres types de risques - les cadres tels que ceux mentionnés ci-dessus peuvent éclairer les considérations relatives à la sécurité et à la protection de la vie privée dans les fonctions **MAP**, **MEASURE** et **MANAGE** du cadre de gestion des risques liés à l'IA.

Dans le même temps, les orientations disponibles avant la publication du présent CMR sur l'IA n'abordent pas de manière exhaustive de nombreux risques liés aux systèmes d'IA. Par exemple, les cadres et orientations existants ne sont pas en mesure de :

- gérer de manière adéquate le problème des préjugés nuisibles dans les systèmes d'IA ;
- faire face aux risques liés à l'IA générative ;
- de traiter de manière exhaustive les problèmes de sécurité liés à l'évasion, à l'extraction de modèles, à l'inférence de membres, à la disponibilité ou à d'autres attaques liées à l'apprentissage automatique ;
- tenir compte de la surface d'attaque complexe des systèmes d'IA ou d'autres abus de sécurité permis par les systèmes d'IA ; et
- prendre en compte les risques associés aux technologies d'IA de tiers, à l'apprentissage par transfert et à l'utilisation hors label, où les systèmes d'IA peuvent être formés pour prendre des décisions en dehors des contrôles de sécurité d'une organisation ou formés dans un domaine, puis "affinés" pour un autre domaine.

Tant l'IA que les technologies et systèmes logiciels traditionnels font l'objet d'innovations rapides. Il convient de suivre et de déployer les avancées technologiques afin de tirer parti de ces développements et d'œuvrer à un avenir de l'IA qui soit à la fois digne de confiance et responsable.

Annexe C :
Gestion des risques liés à l'IA et interaction entre l'homme et l'IA

Les organisations qui conçoivent, développent ou déploient des systèmes d'IA destinés à être utilisés dans des contextes opérationnels peuvent améliorer leur gestion des risques liés à l'IA en comprenant les limites actuelles de l'interaction entre l'homme et l'IA. Le cadre de référence de l'IA permet de définir et de différencier clairement les divers rôles et responsabilités de l'homme lors de l'utilisation, de l'interaction ou de la gestion des systèmes d'IA.

Bon nombre des approches fondées sur les données sur lesquelles s'appuient les systèmes d'IA tentent de convertir ou de représenter les pratiques d'observation et de prise de décision individuelles et sociales en quanti- tés mesurables. La représentation de phénomènes humains complexes à l'aide de modèles mathématiques peut se faire au prix de la suppression du contexte nécessaire. Cette perte de contexte peut à son tour rendre difficile la compréhension des impacts individuels et sociétaux qui sont essentiels aux efforts de gestion des risques de l'IA.

Les questions qui méritent d'être approfondies et étudiées sont les suivantes :

1. **Les rôles et responsabilités de l'homme dans la prise de décision et la supervision des systèmes d'IA doivent être clairement définis et différenciés.** Les configurations entre l'homme et l'IA peuvent aller d'un système entièrement autonome à un système entièrement manuel. Les systèmes d'IA peuvent prendre des décisions de manière autonome, s'en remettre à un expert humain ou être utilisés par un décideur humain en tant qu'avis complémentaire. Certains systèmes d'IA peuvent ne pas nécessiter de supervision humaine, comme les modèles utilisés pour améliorer la compression vidéo. D'autres systèmes peuvent nécessiter spécifiquement une supervision humaine.

2. **Les décisions relatives à la conception, au développement, au déploiement, à l'évaluation et à l'utilisation des systèmes d'IA reflètent des biais cognitifs systémiques et humains.** Les acteurs de l'IA intègrent leurs biais cognitifs, tant individuels que collectifs, dans le processus. Les biais peuvent découler des tâches de prise de décision de l'utilisateur final et être introduits tout au long du cycle de vie de l'IA par le biais d'hypothèses, d'attentes et de décisions humaines lors des tâches de conception et de modélisation. Ces biais, qui ne sont pas nécessairement toujours nuisibles, peuvent être exacerbés par l'opacité du système d'IA et le manque de transparence qui en résulte. Les préjugés systémiques au niveau de l'organisation peuvent influencer la manière dont les équipes sont structurées et qui contrôle les processus décisionnels tout au long du cycle de vie de l'IA. Ces préjugés peuvent également influencer les décisions prises en aval par les utilisateurs finaux, les et les responsables politiques, et avoir des répercussions négatives.

3. **Les résultats de l'interaction entre l'homme et l'IA varient.** Dans certaines conditions - par exemple, dans les tâches de jugement basées sur la perception - la partie IA de l'interaction homme-AI peut amplifier les préjugés humains, conduisant à des décisions plus biaisées que l'IA ou l'homme seul. Toutefois, lorsque ces variations sont judicieusement prises en compte dans l'organisation des équipes homme-AI, elles peuvent entraîner une complémentarité et une amélioration des performances globales.

4. **La présentation des informations des systèmes d'IA aux humains est complexe.** Les humains perçoivent les résultats et les explications des systèmes d'IA et en tirent un sens de différentes manières, en fonction de leurs préférences, de leurs caractéristiques et de leurs compétences individuelles.

La fonction **GOVERN** offre aux organisations la possibilité de clarifier et de définir les rôles et les responsabilités des humains dans les configurations de l'équipe Human-AI et de ceux qui supervisent les performances du système d'IA. La fonction GOUVERNEMENT crée également des mécanismes permettant aux organisations de rendre leurs processus décisionnels plus explicites, afin de lutter contre les préjugés systémiques.

La fonction **MAP** offre des possibilités de définir et de documenter les processus permettant aux opérateurs et aux praticiens de maîtriser les concepts de performance et de fiabilité des systèmes d'IA, et de définir les normes techniques et les certifications pertinentes. La mise en œuvre des catégories et sous-catégories de la fonction **MAP** peut aider les organisations à améliorer leurs compétences internes en matière d'analyse du contexte, d'identification des limites des procédures et des systèmes, d'exploration et d'examen des incidences des systèmes fondés sur l'IA dans le monde réel et d'évaluation des processus décisionnels tout au long du cycle de vie de l'IA.

Les fonctions **GOVERN** et **MAP** décrivent l'importance de l'interdisciplinarité et de la diversité des équipes démographiques, ainsi que de l'utilisation du retour d'information des personnes et des communautés susceptibles d'être touchées. Les acteurs de l'IA mentionnés dans le CMR de l'IA qui effectuent des tâches et des activités liées aux facteurs humains peuvent aider les équipes techniques en ancrant les pratiques de conception et de développement dans les intentions des utilisateurs et des représentants de la communauté de l'IA au sens large, ainsi que dans les valeurs sociétales. Ces acteurs contribuent en outre à intégrer les normes et valeurs spécifiques au contexte dans la conception des systèmes et à évaluer les expériences de l'utilisateur final - en liaison avec les systèmes d'IA.

Les approches de gestion des risques de l'IA pour les configurations homme-IA seront renforcées par des recherches et des évaluations en cours. Par exemple, la mesure dans laquelle les humains sont habilités et incités à remettre en question les résultats des systèmes d'IA doit faire l'objet d'études plus approfondies. Il pourrait être utile de recueillir et d'analyser des données sur la fréquence et la justification avec lesquelles les humains annulent les résultats des systèmes d'IA dans les systèmes déployés.

Annexe D :
Attributs de l'AI RMF

Le NIST a décrit plusieurs attributs clés du cadre de référence pour l'IA lorsque les travaux sur le cadre ont commencé. Ces attributs sont restés intacts et ont été utilisés pour guider le développement du cadre de référence pour l'IA. Ils sont présentés ici à titre de référence.

L'AI RMF s'efforce :

1. Elle doit être fondée sur le risque, efficace en termes de ressources, favorable à l'innovation et volontaire.
2. être fondé sur le consensus et être élaboré et régulièrement mis à jour dans le cadre d'un processus ouvert et trans-parental. Toutes les parties prenantes doivent avoir la possibilité de contribuer à l'élaboration du cadre de référence pour l'IA.
3. Utiliser un langage clair et simple, compréhensible par un large public, y compris les cadres supérieurs, les fonctionnaires, les dirigeants d'organisations non gouvernementales et les personnes qui ne sont pas des professionnels de l'IA, tout en restant suffisamment technique pour être utile aux praticiens. Le cadre de référence de l'IA doit permettre de communiquer les risques liés à l'IA au sein de l'organisation, entre les organisations, avec les clients et avec le grand public.
4. Fournir un langage et une compréhension communs pour gérer les risques liés à l'IA. Le cadre de référence pour l'IA devrait proposer une taxonomie, une terminologie, des définitions, des mesures et des caractérisations des risques liés à l'IA.
5. être facilement utilisable et s'intégrer à d'autres aspects de la gestion des risques. L'utilisation du cadre devrait être intuitive et facilement adaptable dans le cadre de la stratégie et des processus plus larges de gestion des risques d'une organisation. Il doit être cohérent ou aligné sur d'autres approches de la gestion des risques liés à l'IA.
6. être utile à un large éventail de perspectives, de secteurs et de domaines technologiques. Le cadre de référence de l'IA devrait être universellement applicable à toute technologie d'IA et à des cas d'utilisation spécifiques au contexte.
7. Être axé sur les résultats et non prescriptif. Le cadre devrait fournir un catalogue de résultats et d'approches plutôt que de prescrire des exigences uniques.
8. Tirer parti des normes, lignes directrices, meilleures pratiques, méthodologies et outils existants pour la gestion des risques liés à l'IA et les faire mieux connaître, et illustrer la nécessité de ressources supplémentaires et améliorées.
9. Être agnostique en matière de législation et de réglementation. Le cadre doit permettre aux organisations d'opérer dans le cadre des régimes juridiques ou réglementaires nationaux et internationaux applicables.
10. Être un document évolutif. Le cadre de gestion des risques liés à l'IA devrait être facilement mis à jour en fonction de l'évolution des technologies, des connaissances et des approches en matière de fiabilité et d'utilisation de l'IA, ainsi que des enseignements tirés par les parties prenantes de la mise en œuvre de la gestion des risques liés à l'IA en général et du présent cadre en particulier.

Cette publication est disponible gratuitement à
l'adresse suivante :
https://doi.org/10.6028/NIST.AI.100-1