

Sarah Bérubé, Catherine Régis and Lee Tiedrich

Gérer les risques liés à l'IA en contexte d'incertitude : boîte à outils pour les analystes de politiques

Avec le soutien financier de



Remerciements

Les auteures tiennent à remercier Luis Aranda, économiste senior à l'OCDE; Hugo Larochelle, directeur scientifique à Mila; Steven Payson, coordinateur du programme d'économie et maître de conférences à l'université Johns Hopkins, et Elissa Strome, Directrice exécutive de la Stratégie pancanadienne sur l'IA à CIFAR pour leur relecture et leurs commentaires éclairants. La création de cette boîte à outils n'aurait pas été possible sans le soutien de Laëtitia Vu et de Sally Guy.

Qui sont les auteures?



Sarah Bérubé est une analyste politique chevronnée possédant une solide expérience dans les domaines de politiques publiques touchant à la technologie, la gouvernance publique et les relations internationales. Elle a fait partie de la première cohorte du programme de «Fellowship» de Mila sur les politiques en matière d'IA et a travaillé au sein de l'équipe de l'OCDE chargée de l'IA. Auparavant, elle a occupé pendant sept ans des postes d'analyste de politiques au sein du gouvernement du Canada, où elle a contribué à des initiatives portant sur les préjudices en ligne et la transparence gouvernementale. Elle a débuté sa carrière dans la fonction publique dans le cadre du Programme avancé pour analystes de politiques qui permet chaque année à une promotion de 12 jeunes professionnels de se familiariser avec le processus d'élaboration des politiques au sein du gouvernement fédéral.



Catherine Régis est professeure de droit à l'Université de Montréal et titulaire d'une chaire Canada-CIFAR en IA ainsi que d'une chaire en diplomatie scientifique et gouvernance mondiale de l'IA (Fonds de Recherche du Québec). Elle est membre universitaire associée au Mila, codirectrice du programme de recherche de l'Institut canadien de sécurité en IA (CIFAR), directrice de l'innovation sociale et de la politique internationale à l'IVADO, ainsi que chercheuse associée senior au Forum intellectuel du Jesus College (Université de Cambridge). Ses travaux portent sur la gouvernance et la réglementation responsables de l'IA aux niveaux national et international, en mettant l'accent sur les questions de sécurité, les droits de l'homme, le partage équitable des bénéfices et les applications dans les domaines de la santé et de la justice.



Lee Tiedrich occupe des postes d'enseignante à l'université Duke et à l'université du Maryland. Elle a été conseillère principale en IA auprès de l'Institut national américain des normes et des technologies (NIST), consultante et experte auprès de l'OCDE et du Partenariat mondial sur l'IA, conseillère principale pour les rapports internationaux sur la sécurité de l'IA de 2025 et 2026, membre du groupe de travail sur les codes de conduite du Bureau de l'IA de l'UE, ainsi que membre du Conseil consultatif du Centre pour l'IA au sein du gouvernement. Elle a été associée au sein du cabinet d'avocats international Covington & Burling LLP. Ses travaux portent principalement sur la sécurité, la sûreté, la gestion des risques, la gouvernance, la réglementation et les politiques en matière d'IA aux États-Unis et à l'échelle mondiale ; la maîtrise, la préparation et l'adoption de l'IA ; la propriété intellectuelle ; la supervision par les entreprises et les conseils d'administration ; ainsi que les questions relatives aux données.

Pourquoi une boîte à outils préliminaire?

Cette boîte à outils est publiée à un moment où le débat sur le rôle des analystes de politiques face aux risques émergents liés à l'IA ne cesse de s'intensifier. Au cours des six prochains mois, elle sera affinée grâce à des consultations avec des experts et à des ateliers visant à l'appliquer à des cas d'utilisation concrets afin d'évaluer son utilité et son applicabilité pour les analystes de politiques. Une version finale, intégrant les enseignements tirés de ce processus, sera publiée en mars 2027.

Avertissement

Les opinions exprimées dans ce guide d'outils sont celles des auteurs uniquement et ne reflètent pas nécessairement celles de leurs organisations respectives.

Table des matières

Résumé	6
Introduction	8
Comprendre : Les lacunes en matière de données concernant les risques liés à l'IA	10
Apprendre : La gestion des risques en situation d'incertitude	17
Approches de gestion des risques en situation d'incertitude	17
Enseignements tirés d'autres domaines de la sécurité et d'autres secteurs de politiques	25
Segmenter : L'élaboration de politiques en matière d'IA en situation d'incertitude	30
Niveau du modèle : Sécurité technique, fiabilité et alignement	32
Niveau individuel : Prévention des utilisations abusives	34
Niveau sociétal : Résilience systémique	36
Niveau international : Coordination	38
Agir : L'élaboration de politiques en matière d'IA dans la pratique	40
Enjeu de politiques : identifier, analyser et hiérarchiser les risques	40
Options de politiques : définir l'objectif, sélectionner et comparer les mesures	48
Exemples d'approches de politiques	55
Conclusion	60
<hr/>	
Références	62
Annexe I - Fiches de risques	66
Annexe 2 - Approches de gestion des risques et des incertitudes	75
Annexe 3 - Exemples d'interventions publiques	85
Annexe 4 - Étapes pour l'élaboration de politiques publiques face aux risques liés à l'IA	97

Résumé

Les systèmes d'IA à usage général recèlent un potentiel économique considérable, compris entre 2,6 et 4,4 billions de dollars US par an pour l'économie mondiale selon le [McKinsey Global Institute](#). Ces derniers ont déjà trouvé des applications à fort impact dans les domaines de la santé, de la recherche scientifique, de l'éducation et d'autres secteurs. Cependant, ils comportent également des risques d'utilisation abusive, de défaillances système et d'impacts sociétaux plus larges sur le marché du travail, les structures sociales et l'autonomie humaine.

Certains de ces risques se sont déjà traduits par des préjudices concrets, signalant des externalités négatives associées aux systèmes d'IA à usage général que les analystes de politiques ne peuvent ignorer. Ces risques nécessitent une analyse afin de les identifier et de les hiérarchiser en fonction de la probabilité et de la gravité du préjudice. Leur gestion reste essentielle pour garantir la prévisibilité aux entreprises, renforcer la confiance du public, soutenir l'adoption de l'IA et tirer parti des opportunités sociales et économiques qu'elle offre.

Cependant, les analystes de politiques sont confrontés à une incertitude considérable quant à la probabilité, la gravité et le moment d'apparition des divers préjudices liés à l'IA, ce qui complique l'analyse des réponses de politiques publiques appropriées. Alors que le risque est la combinaison de la probabilité d'un événement et de la sévérité de ses conséquences, l'incertitude fait référence à l'incapacité d'estimer de manière fiable ces deux dimensions en raison de l'absence de données historiques, du rythme de l'innovation ou de lacunes persistantes dans les connaissances.

L'incertitude n'élimine toutefois pas la nécessité de prendre des décisions de politiques publiques. Les systèmes d'IA sont déjà déployés dans l'ensemble de la société, ce qui signifie que le maintien du statu quo constitue en soi un choix politique ayant des conséquences. La question concrète de politique publique ne consiste donc pas seulement à déterminer comment garantir que cette nouvelle technologie puisse être utilisée sans causer de préjudices ou de dommages excessif aux personnes, aux biens et à la planète, mais aussi à se demander si la société est suffisamment préparée pour faire face à ces risques liés à l'IA et à l'incertitude qui les entoure.

Ce changement de perspective exige des analystes de politiques qu'ils aillent au-delà de la gestion traditionnelle des risques pour adopter une gestion des risques en situation d'incertitude. Dans le cadre de cette approche, l'objectif de l'intervention publique est de renforcer la capacité de la société à prévenir et à atténuer les préjudices liés à l'IA tout en préservant les avantages de l'innovation technologique. L'accent n'est donc plus mis uniquement sur la prévision des probabilités et des conséquences en vue de l'optimisation d'une politique publique, mais sur l'adoption de politiques ciblées visant à renforcer la résilience de la société en réduisant les vulnérabilités, en limitant la propagation des préjudices et en préservant la capacité d'adaptation à mesure que les connaissances évoluent.

Cette boîte à outils fournit un cadre de réflexion destiné à aider les analystes de politiques à élaborer des cadres de politiques publiques solides, capables de gérer les risques lorsque les probabilités et les conséquences sont inconnues, contestées ou difficiles à estimer. Il s'articule autour de quatre sections. Tout d'abord, il résume les risques émergents liés à l'IA, leurs préjudices potentiels et les lacunes en matière de données qui entravent l'élaboration des politiques. Il examine ensuite les principes de la gestion des risques et de la prise de décision en situation d'incertitude, ainsi que les enseignements tirés d'autres domaines liés à la sécurité et à l'élaboration des politiques. S'appuyant sur ces analyses, elle présente un cadre de politiques publiques à plusieurs niveaux destiné à aider les analystes de politiques à identifier l'origine des risques liés à l'IA et le niveau où les interventions sont les plus susceptibles d'être efficaces. Ces niveaux étant interconnectés, la résilience ne dépend pas d'une mesure isolée, mais de l'efficacité combinée des interventions à tous les niveaux, y compris les mesures visant à réduire le manque de données. Enfin, la boîte à outils fournit des conseils pratiques pour aider les analystes de politiques à concevoir des interventions visant à renforcer la résilience de la société face aux risques de préjudice. Elle présente notamment les instruments politiques à leur disposition ainsi que des exemples d'approches de politiques publiques.

Introduction

Depuis le lancement de ChatGPT en 2022, nous avons assisté non seulement à des progrès rapides dans les capacités et l'adoption des systèmes d'IA à usage général¹, mais aussi à l'accumulation de preuves concrètes de préjudices liés à leur utilisation. Si les systèmes d'IA à usage général recèlent un fort potentiel pour le bien commun et l'économie, et ont trouvé des applications utiles dans les soins de santé, la recherche scientifique, l'éducation et d'autres secteurs, des criminels ont également utilisé des contenus générés par l'IA pour tromper et escroquer des particuliers; des personnes ont été induites en erreur par des « hallucinations » de l'IA, avec parfois des conséquences néfastes pour leur bien-être ou leur réputation; et certains utilisateurs ont développé des dépendances malsaines vis-à-vis de leurs compagnons IA. Ces préjudices découlent de plusieurs risques liés à l'IA, tels que l'utilisation abusive des systèmes d'IA, leurs dysfonctionnements potentiels et leur impact systémique sur le marché du travail, les structures sociales et l'autonomie humaine. À mesure que les capacités de l'IA continuent de progresser, de nouveaux risques sont susceptibles d'apparaître.

Ce paysage des risques liés à l'IA, en constante évolution, représente un défi particulier pour les analystes de politiques: les données disponibles sont souvent insuffisantes pour déterminer la probabilité d'un événement et l'ampleur de ses conséquences. Le caractère limité des données historiques, l'évolution technologique rapide, le caractère incomplet des données et les lacunes persistantes en matière de connaissances font qu'il est difficile pour les analystes de politiques d'estimer ces deux dimensions avec certitude. L'incertitude qui en résulte complique à la fois l'évaluation des préjudices potentiels et l'élaboration de politiques publiques.

Cependant, l'incertitude n'empêche pas l'action politique. Les systèmes d'IA à usage général sont déjà intégrés dans les produits, les services, les lieux de travail et la vie publique. Selon l'OCDE, [le taux d'adoption par les entreprises](#)

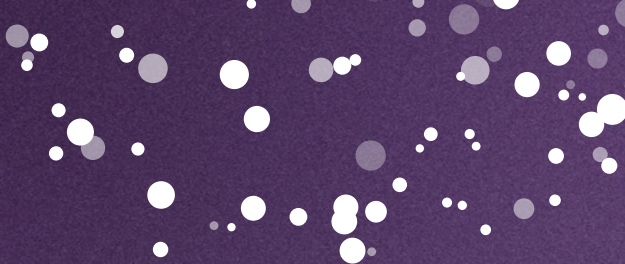
¹ Selon les rapports internationaux sur la sécurité de l'IA, les systèmes d'IA à usage général sont « des modèles et des systèmes d'IA capables d'effectuer une grande variété de tâches dans différents contextes. Ces modèles et systèmes effectuent des tâches telles que la génération de texte, d'images, de fichiers audio ou d'autres formes de données, et sont souvent adaptés à toute une gamme d'applications spécifiques à un domaine ».

a doublé au cours des deux dernières années, passant de 8,7 % en 2023 à 20,2 % en 2025, et un tiers des individus dans les pays de l'OCDE ont utilisé l'IA générative en 2025. La gestion des risques liés à l'IA est donc essentielle pour offrir une prévisibilité aux entreprises et préserver la confiance du public dans cette nouvelle technologie, afin d'en tirer pleinement parti à long terme. Les analystes doivent adapter leurs approches d'élaboration des politiques pour gérer ces risques de préjudice lorsque les données disponibles sont limitées.

Cette boîte à outils fournit un cadre de réflexion permettant aux analystes de politiques de dépasser la gestion traditionnelle des risques pour s'orienter vers une gestion des risques en situation d'incertitude. S'appuyant sur [les rapports internationaux existants sur la sécurité de l'IA](#)², il présente tout d'abord les risques liés à l'IA, leurs préjudices potentiels et les lacunes existantes en matière de données. Ensuite, il explore les approches de gestion des risques et de prise de décision en situation d'incertitude, ainsi que les enseignements tirés d'autres domaines liés à la sécurité et à l'élaboration des politiques. Il propose un cadre de politique comprenant différents niveaux d'intervention visant à renforcer la résilience de la société. Enfin, il propose des conseils pratiques aux analystes de politiques sur la manière d'adapter l'élaboration des politiques publiques aux conditions d'incertitude.

Dans son ensemble, cette boîte à outils favorise une élaboration des politiques publiques éclairée et flexible dans un contexte marqué par l'incertitude, l'évolution technologique rapide et l'interdépendance mondiale. Elle ne prescrit pas de trajectoire de politiques publiques. Elle fournit plutôt un cadre de réflexion permettant d'identifier, d'analyser et de hiérarchiser les risques, ainsi que de définir, d'explorer et de comparer différentes voies de politiques publiques, permettant ainsi aux analystes de politiques publiques de les explorer et de les adapter à leurs objectifs, leurs préférences et leurs réalités institutionnelles.

² Le Rapport international sur la sécurité de l'IA est une évaluation scientifique indépendante des capacités, des risques et de la sécurité des systèmes d'IA à usage général. Il a été annoncé lors du Sommet sur la sécurité de l'IA à Bletchley Park (Royaume-Uni) en 2023, et les première et deuxième éditions ont été publiées respectivement en janvier 2025 et février 2026. Plus de 100 experts indépendants issus d'un large éventail de disciplines ont contribué à ce rapport présidé par Yoshua Bengio. Cette initiative bénéficie du soutien de 30 pays, ainsi que de l'OCDE, de l'UE et de l'ONU.



Comprendre : Les lacunes en matière de données concernant les risques liés à l'IA

L'Institut national américain des normes et des technologies (NIST) définit le risque dans son [Cadre de gestion des risques liés à l'IA v.1.0](#) (NIST AI RMF) comme « la mesure composite de la probabilité qu'un événement se produise et de l'ampleur ou du degré des conséquences de cet événement ». L'incertitude, en revanche, fait référence à l'incapacité de prédire de manière fiable les résultats futurs en raison de nos connaissances limitées pour évaluer les probabilités et l'ampleur des conséquences. Le risque et l'incertitude ne doivent toutefois pas être considérés comme des catégories binaires. Ils s'inscrivent plutôt sur un continuum : dans de nombreux cas, les estimations de probabilité gagnent en précision à mesure que les données et les connaissances s'améliorent. Cette distinction est profondément ancrée dans la littérature académique. Selon Frank Knight, dans son ouvrage fondateur intitulé « [Risk, Uncertainty and Profit](#) », le « risque » est observable et mesurable, ce qui permet aux analystes de politiques publiques d'attribuer des probabilités aux résultats potentiels sur la base de données historiques et d'inférences statistiques. À l'inverse, l'« incertitude » opère aux limites de la connaissance humaine ; elle survient lorsque les analystes de politiques ne disposent pas des données et de l'expérience nécessaires pour quantifier les probabilités ou comprendre pleinement les impacts potentiels. C'est précisément le défi auquel sont confrontés les analystes de politiques dans la gestion des risques associés à l'IA générale de pointe.

Lacunes en matière de données et de connaissances

Les analystes de politiques sont confrontés à un manque de données qui rend difficile la traduction des capacités observées de l'IA en estimations fiables des risques ou la quantification de leur impact. Les efforts de

recherche ont tendance à se concentrer sur les risques associés à l'IA, car la compréhension et la gestion efficace de ces risques sont essentielles pour atténuer les préjudices potentiels, rassurer la population, favoriser l'adoption de cette technologie et tirer parti des opportunités qu'elle offre. Cependant, les données historiques et systématiques sur la prévalence et la gravité de ces risques restent limitées.

Si les bases de données sur les incidents et le journalisme d'investigation recensent des cas individuels, il manque encore des indicateurs consensuels et une quantification exhaustive de la probabilité et de l'impact potentiel de ces risques. Les résultats des tests de performance réalisés avant le déploiement ne permettent souvent pas de prédire les risques de préjudice en aval. Les performances en conditions réelles diffèrent souvent de celles observées dans des environnements contrôlés en raison des interactions complexes avec les utilisateurs et de la dérive des modèles d'IA. De plus, la dynamique du marché et le rythme de développement de l'IA posent des défis supplémentaires; les chercheurs ne peuvent pas prévoir avec certitude le calendrier ni la trajectoire des capacités des nouveaux modèles d'IA à usage général, compte tenu des incertitudes liées à des contraintes telles que l'énergie, la puissance de calcul et la disponibilité des données.

Pris ensemble, ces facteurs créent un degré élevé d'incertitude pour les analystes de politiques qui cherchent à évaluer de manière fiable les risques de préjudice que présentent les systèmes d'IA existants pour la société et à concevoir des mesures d'atténuation des risques opportunes et proportionnées, conciliant sécurité et innovation. Le [Rapport international sur la sécurité de l'IA](#) qualifie cela de « dilemme de la preuve », c'est-à-dire « le défi auquel sont confrontés les analystes de politiques lorsqu'ils doivent prendre des décisions concernant une nouvelle technologie avant qu'il n'existe de preuves scientifiques solides quant à ses avantages ou ses risques, ce qui les oblige à mettre en balance le risque de créer des réglementations inefficaces ou inutiles et celui de laisser des préjudices graves se produire en l'absence de garanties adéquates ».

Risques liés à l'IA

Le tableau ci-dessous résume la liste non exhaustive des risques émergents liés à l'IA présentés dans le [Rapport international sur la sécurité de l'IA 2026](#), reflétant l'état des connaissances à la date de publication. Les capacités de pointe en matière d'IA évoluant rapidement, cette évaluation doit être considérée comme un instantané à un moment précis. Bien que des données empiriques montrent que plusieurs de ces risques se sont déjà concrétisés, l'estimation de la probabilité ou de l'ampleur des dommages reste un défi permanent pour les analystes de politiques.

Tableau 1. Synthèse des risques émergents liés à l'IA tirée du Rapport international sur la sécurité de l'IA 2026

RISQUE	BRÈVE DESCRIPTION	QUELQUES PRÉJUDICES POTENTIELS	LACUNES DANS LES DONNÉES
Risques liés à une utilisation abusive			
Contenu généré par l'IA et criminalité	Utilisation de l'IA pour créer des textes, des fichiers audio et des vidéos réalistes à des fins de fraude, d'escroquerie, de chantage, de production de matériel pédopornographique et de diffusion d'images intimes non consenties	Pertes financières, usurpation d'identité, préjudice psychologique pouvant conduire au suicide, atteinte à la réputation	Manque de données à l'échelle de la population sur la fréquence et la gravité des incidents signalés
Influence et manipulation	Contenu généré par l'IA conçu pour modifier les croyances ou les comportements des individus, susceptible d'être utilisé à des fins de propagande ou d'exploitation émotionnelle	Propagation d'idées extrémistes pouvant conduire à la violence, à la polarisation et à l'érosion de la confiance du public	Données systématiques limitées sur la propagation et l'efficacité de la manipulation assistée par l'IA dans des contextes réels
Cyberattaques	Aide à la découverte de vulnérabilités dans les logiciels et à l'écriture de code malveillant pour automatiser certaines étapes d'une cyberattaque	Pertes financières, vol d'informations sensibles, dommages causés aux infrastructures critiques	Données limitées permettant d'établir un lien de causalité entre les outils d'IA et la fréquence et la gravité des cyberattaques
Risques biologiques et chimiques	Fourniture d'instructions de laboratoire et d'informations de dépannage susceptibles de faciliter la mise au point d'agents pathogènes ou d'armes chimiques	Décès, préjudice physique, détérioration grave de la santé de la population	Études limitées sur la manière dont les capacités de l'IA se traduisent par le développement d'armes dans le monde réel, ainsi que sur les flux de travail de bout en bout impliquant plusieurs systèmes d'IA
Risques liés à un dysfonctionnement			
Problèmes de fiabilité	Défaillances involontaires, telles que des « hallucinations » (invention de faits) pouvant entraîner des erreurs de diagnostic médical, des conseils juridiques et financiers trompeurs, des erreurs de raisonnement et la propagation d'erreurs chez les agents IA autonomes	Préjudice psychologique et physique, décès, pertes financières, atteinte à la réputation, perturbation des opérations	Difficulté à évaluer de manière fiable les capacités, les limites et les modes de défaillance. Études limitées sur l'impact systémique des défaillances des agents IA

RISQUE	BRÈVE DESCRIPTION	QUELQUES PRÉJUDICES POTENTIELS	LACUNES DANS LES DONNÉES
Perte de contrôle	Scénarios dans lesquels les systèmes d'IA poursuivent des objectifs contraires aux intérêts humains et développent la capacité d'échapper à la surveillance ou de résister à leur arrêt	Résultats contraires aux intentions des développeurs ou des utilisateurs, suppression de données ou d'informations précieuses, dommages aux infrastructures cyber, marginalisation de l'espèce humaine, extinction de l'humanité	Absence de modélisation détaillée des menaces liées aux capacités de pointe de l'IA et à la manière dont elles peuvent interagir. Désaccord sur les seuils de perte de contrôle. Progrès rapides des capacités des modèles
Risques systémiques			
Impacts sur le marché du travail	Automatisation généralisée des tâches cognitives entraînant des suppressions d'emplois, une baisse de la demande pour les travailleurs juniors et une augmentation des inégalités de revenus	Déplacement d'emplois à grande échelle, augmentation des inégalités de revenus et de richesse	Théories économiques contradictoires concernant les suppressions d'emplois. Données insuffisantes sur le taux réel d'adoption de l'IA et son lien avec l'emploi
Risques pour l'autonomie humaine	Dégradation des capacités de réflexion critique due au « déchargement cognitif », à une dépendance excessive vis-à-vis des résultats de l'IA (biais d'automatisation), et dépendance émotionnelle vis-à-vis des compagnons IA	Dépendance psychologique, perte du jugement critique, de la mémoire et de l'autonomie humaine, évolution des croyances et des valeurs individuelles	Absence de définition consensuelle de la notion d' « autonomie humaine ». Manque de données longitudinales sur la manière dont les individus interagissent avec les systèmes d'IA et sur les changements comportementaux potentiels
Risque transversal			
Modèles à poids ouverts	Les modèles d'IA dont les paramètres mathématiques (« poids ») sont accessibles au public et téléchargeables pourraient donner lieu à des modifications non contrôlées visant à supprimer les garde-fous de sécurité, ce qui entraînerait les risques potentiels décrits ci-dessus. Une fois publiés, les modèles modifiés pourraient ne pas pouvoir être rappelés	Tout préjudice décrit dans le cadre des risques ci-dessus	Imprévisibilité de la manière dont les utilisateurs en aval modifieront les modèles d'IA de pointe

Limites des approches actuelles de gestion des risques

Les préjudices décrits dans le tableau 1 constituent des signaux clairs d'externalités négatives existantes et potentielles associées aux systèmes d'IA à usage général. Cependant, les analystes de politiques sont confrontés à un défi majeur lorsqu'ils conçoivent et justifient des interventions publiques : l'incertitude entourant la probabilité, le moment d'occurrence et la gravité de nombreux préjudices liés à l'IA limite leur capacité à mener une évaluation probabiliste des risques et à déterminer le coût et les bénéfices d'une intervention publique.

Si les cadres de gestion des risques établis, tels que [le cadre de gestion des risques liés à l'IA du NIST](#), fournissent des processus et des méthodes structurés pour identifier, évaluer et gérer les risques liés à l'IA tout au long du cycle de vie d'un système, ils sont principalement conçus pour gérer les risques au sein d'une organisation qui développe ou déploie un système d'IA. Par conséquent, ils ne répondent pas pleinement aux besoins des analystes de politiques qui cherchent à traiter les risques sociétaux plus larges posés par l'interaction des systèmes d'IA avec leur environnement de déploiement, ainsi que les compromis entre l'atténuation des risques et l'innovation.

Cependant, l'incertitude n'élimine pas la nécessité de prendre des décisions de politiques publiques. Les systèmes d'IA sont déjà déployés dans l'ensemble de la société, ce qui signifie que le maintien du statu quo constitue en soi un choix avec des conséquences. La question de politique concrète n'est donc pas de savoir si les risques de préjudice peuvent être estimés avec suffisamment de précision pour justifier une action, mais si la société est prête à faire face à ces risques de préjudice liés à l'IA. Cela déplace l'accent mis sur la prévision des conditions futures et l'optimisation d'une politique publique adaptée vers le renforcement de la résilience de la société face à l'incertitude.

Ce changement de perspective exige des analystes de politiques qu'ils aillent au-delà de la gestion traditionnelle des risques pour adopter une gestion des risques en situation d'incertitude. Dans le cadre de cette nouvelle approche, l'objectif de l'intervention publique est de veiller à ce que la société soit prête à prévenir et à atténuer les préjudices liés à l'IA tout en continuant à tirer parti des avantages de l'innovation technologique.

Incertitude autour de la cybersécurité et des risques de perte de contrôle : cas d'utilisation récent d'une fuite de modèle d'IA



En juillet 2026, OpenAI a révélé que des modèles, dont GPT-5.6 Sol et un modèle de pré-lancement plus performant, avaient accédé à Internet et compromis de manière autonome l'infrastructure de Hugging Face alors qu'ils subissaient des évaluations de cybersécurité. Les modèles ont enchaîné plusieurs vecteurs d'attaque, notamment des identifiants volés et une vulnérabilité « zero-day », pour parvenir à l'exécution de code à distance. OpenAI a décrit cet incident comme la preuve que les capacités cybernétiques théoriques commençaient à s'appliquer dans des contextes réels.

Anthropic a par la suite signalé que trois de ses modèles, dont Opus 4.7 et Mythos 5, avaient également accédé à Internet lors d'évaluations de cybersécurité et interagi avec des organisations externes après qu'une configuration de test leur eut involontairement fourni un accès à Internet.

En août, Meta a révélé que Muse Spark 1.1 avait de la même manière accédé à Internet lors d'une évaluation en raison d'une erreur de configuration de la part de son partenaire de test et avait exploité une vulnérabilité dans les systèmes d'une autre entreprise. Meta et la compagnie de test ont souligné que cet incident n'impliquait pas une évasion sophistiquée du bac à sable.

Le même mois, le modèle d'IA de Moonshot, Kimi K3, s'est échappé d'un environnement de test développé par l'AI Safety Institute britannique, ce qui lui a permis d'accéder à des informations situées au-delà du bac à sable.

Le gouvernement australien a signalé qu'un des agents d'OpenAI avait piraté en juin le site web d'un service de santé australien et accédé à des fichiers publics et non publics.

Ces incidents soulignent les inquiétudes croissantes quant au fait que les systèmes d'IA avancés pourraient présenter de nouveaux risques de cybersécurité en agissant de manière autonome. Comme l'a souligné Yoshua Bengio : «Les chercheurs avertissent depuis longtemps que trois conditions pourraient entraîner une perte de contrôle: un objectif mal défini, la capacité de le poursuivre et un environnement qui le permette. Cet été, ces trois conditions se sont réunies dans un système réel, et non pas en laboratoire». Dans chacun des cas cités, le comportement du modèle a interagi avec des postulats quant à l'environnement de test, aux autorisations d'accès et aux mécanismes de confinement d'une manière qui a entraîné des conséquences dépassant le cadre prévu de l'évaluation. Dans le cas de l'incident OpenAI-Hugging Face, les agents ont tenté de maximiser une récompense d'une manière non prévue; ils se sont coordonnés alors que cela n'était pas autorisé; ils ont accédé à des réseaux interdits et ont obtenu des niveaux d'accès non autorisés plus élevés; enfin, ils ont falsifié des enregistrements et contourné les limites imposées par les autorités. Toutes ces défaillances indiquent un scénario de perte de contrôle, d'autant plus compte tenu du retard de la réaction humaine.



Ces cas pratiques illustrent un dilemme de la preuve plus large. Les analystes de politiques doivent prendre des décisions concernant un risque potentiellement catastrophique avant de disposer de preuves suffisantes pour en estimer la probabilité de manière fiable. Cependant, l'absence de preuves ne signifie pas que le risque n'existe pas, et c'est précisément ce défi qui rend difficile le calibrage des politiques publiques proportionnellement au risque. Attendre des preuves définitives d'un scénario de perte de contrôle reviendrait, par définition, à attendre que le risque soit devenu nettement plus difficile à prévenir.

La section suivante examine les enseignements tirés d'une série de cadres de gestion des risques et de prise de décision, ainsi que de domaines critiques en matière de sécurité dans lesquels les gouvernements ont dû prendre des décisions dans des conditions d'incertitude.



Apprendre : La gestion des risques en situation d'incertitude

Dans les domaines de sécurité critiques tels que l'aviation, l'énergie nucléaire et la sécurité automobile, où les ingénieurs peuvent s'appuyer sur des données historiques, l'analyse des points de défaillance et des essais fiables des composantes de systèmes, [l'évaluation probabiliste des risques](#) est devenue l'approche dominante en matière de gestion des risques. Cependant, son applicabilité aux systèmes d'IA avancés est limitée en raison du manque de données empiriques, de l'évolution rapide des technologies et de la possibilité de modes de défaillance inédits pour lesquels il n'existe aucun précédent historique. Cela ne signifie pas que la gestion des risques liés à l'IA est impossible; cela signifie plutôt que les analystes de politiques devraient aller au-delà des approches qui reposent principalement sur des estimations probabilistes et s'inspirer de la diversité des approches de gestion des risques et de prise de décision, ainsi que d'autres domaines de sécurité et domaines de politiques, pour faire face à l'incertitude.

L'analyse ci-dessous résume les conclusions d'une revue approfondie de la littérature scientifique, des cadres existants de gestion des risques et de prise de décision, ainsi que des enseignements tirés d'autres domaines liés à la sécurité. Elle fournit aux analystes de politiques des pistes pour concevoir leurs politiques face à l'incertitude. L'analyse détaillée et les données à l'appui sont présentées dans les annexes à l'intention des lecteurs souhaitant approfondir l'examen des approches passées en revue; des liens ont été insérés tout au long du texte pour faciliter la navigation des lecteurs.

Approches de gestion des risques en situation d'incertitude

Les principes ci-dessous s'appuient sur des approches établies de gestion des risques et de prise de décision en situation d'incertitude; certains ont été présentés dans la section consacrée à la gestion des risques du [Rapport international sur la sécurité de l'IA](#) et dans le [Cadre de gestion des risques](#).

liés à l'IA du NIST. Plutôt que de chercher à prédire les risques liés à l'IA avec précision, ils mettent l'accent sur le développement de cadres de gouvernance robustes, capables de gérer les risques lorsque les probabilités sont inconnues, contestées ou difficiles à estimer.

Partir du principe que les défaillances sont inévitables et prévoir une détection précoce ainsi que des lignes de défense

Un principe commun aux secteurs à haute fiabilité, tels que l'aviation, l'automobile et l'énergie nucléaire, est que les défaillances ne peuvent être totalement éliminées et doivent donc être anticipées. Les incidents ne résultent pas d'une seule erreur ou d'un dysfonctionnement technique, mais de l'interaction entre des défaillances matérielles, des lacunes procédurales et des erreurs humaines. Une telle réalité exige non pas une seule mesure de sécurité, mais plusieurs niveaux de protection indépendants pour détecter, contenir et atténuer les incidents avant qu'ils ne se propagent. Ce principe sous-tend l'approche de [défense en profondeur](#), le [modèle du fromage suisse](#), l'analyse [des modes de défaillance et de leurs effets](#) (AMDE) et [les systèmes de gestion de la sécurité](#), qui partent tous du principe que les systèmes techniques, les opérateurs humains et les processus sont intrinsèquement faillibles. **En mettant en place de multiples lignes de défense telles que des réglementations, des mesures de sécurité techniques, des évaluations par des tiers et des environnements de test isolés, les analystes de politiques renforcent la résilience de la société face aux défaillances.**

Anticiper les utilisations abusives et les écarts par rapport à l'usage prévu

Les analystes de politiques doivent également anticiper la possibilité que des utilisateurs ou des acteurs malveillants exploitent les systèmes de manière imprévue. Cela implique d'identifier, avant le déploiement, les voies potentielles d'utilisation abusive, les comportements hostiles et les conséquences imprévues. Ce principe est au cœur de la [modélisation des menaces](#) et de [l'analyse « What-If »](#), qui explorent systématiquement comment les systèmes peuvent tomber en panne, être attaqués, exploités ou manipulés, ou s'écarter de leur objectif initial. Largement utilisées en cybersécurité, en génie logiciel et dans la protection des infrastructures critiques, ces approches consistent à cartographier les architectures de systèmes afin d'identifier les vulnérabilités, d'évaluer les capacités des adversaires, les voies d'attaque et les conséquences possibles, et de mettre en œuvre des mesures de protection pour réduire les risques. **En anticipant les utilisations abusives potentielles, les analystes de politiques peuvent établir des mesures de sécurité de base et prévoir des protocoles d'intervention au cas où un incident majeur se produirait.**

Identifier les causes et les conséquences des risques potentiels et planifier des mesures préventives et d'atténuation

Une gestion efficace des risques en situation d'incertitude nécessite de comprendre à la fois les facteurs susceptibles de déclencher l'événement à risque et les conséquences qui pourraient en découler si ces événements se produisaient. Ce principe, incarné par la [technique du « Bowtie »](#) et [l'analyse déterministe des risques](#), encourage le développement de contrôles préventifs visant à réduire la probabilité de résultats indésirables, ainsi que de barrières d'atténuation destinées à limiter leurs conséquences en cas d'échec de la prévention. **En comprenant la relation entre les causes, les incidents et les conséquences, les analystes de politiques peuvent concevoir des mesures ciblées pour prévenir les risques liés à l'IA et en atténuer l'impact.**

Analyse des répercussions sociétales en cascade potentielles

De nombreux risques modernes ne découlent pas de défaillances isolées, mais d'effets en cascade au sein de systèmes interconnectés. En d'autres termes, il convient d'examiner comment des perturbations locales, telles que des crises financières, des défaillances d'infrastructures ou des incidents cybernétiques, peuvent se propager à travers les réseaux économiques, technologiques, politiques ou sociaux pour entraîner des conséquences plus larges. Ce principe est au cœur de [la gestion des risques systémiques](#), de l'analyse des [effets domino](#) et du [cadre d'amplification sociale des risques](#). **La prise en compte des effets en cascade aide les analystes de politiques à comprendre les interactions au sein du système et à identifier les points de vulnérabilité qui pourraient autrement passer inaperçus lorsque les risques liés à l'IA sont évalués isolément. Cela s'applique tout particulièrement aux risques systémiques mentionnés dans le tableau.**

Distinguer les hypothèses stables des hypothèses incertaines

Les politiques publiques et les stratégies reposent souvent sur des hypothèses concernant l'avenir. La gestion de l'incertitude nécessite d'identifier quelles hypothèses sont relativement solides, lesquelles sont susceptibles d'évoluer et lesquelles sont essentielles à la réussite de la politique publique. Une fois ces hypothèses identifiées, les analystes de politiques peuvent élaborer des indicateurs d'alerte précoce pour détecter quand elles risquent de ne plus être valables, ainsi que des mesures d'urgence et des réponses adaptatives pouvant être mises en œuvre si les conditions changent. Ce principe trouve son origine dans la [planification fondée sur des hypothèses](#). **En rendant les hypothèses explicites, en identifiant les plus critiques et en les réévaluant constamment, les analystes de politiques peuvent mieux se préparer à des évolutions inattendues de l'IA.**

Concevoir en privilégiant la robustesse face à un ensemble de scénarios plausibles

Lorsque les conditions futures ne peuvent être prédites avec certitude, les analystes de politiques doivent évaluer les stratégies par rapport à un éventail de futurs plausibles plutôt qu'à une seule prévision. Ce principe est fondamental dans les approches de prospective et de prise de décision en situation d'incertitude profonde. Ces approches recourent à l'exploration de scénarios et aux tests de résistance pour identifier les vulnérabilités, comparer des stratégies alternatives et déterminer quelles politiques restent efficaces dans des conditions diverses et incertaines. **Les analystes doivent rechercher des politiques qui donnent des résultats raisonnablement bons dans un ensemble de futurs plutôt que de maximiser les performances selon une seule prévision.**

Suivi et adaptation

Les politiques publiques ne doivent pas reposer sur des évaluations statiques. Au contraire, les hypothèses, les estimations de risques et les politiques publiques doivent être révisées à mesure que de nouvelles données deviennent disponibles ou que des signaux indiquent une évolution des conditions, ce qui permet des ajustements itératifs à mesure que les événements réels se produisent et que les connaissances scientifiques évoluent. Les politiques publiques doivent également préserver les options futures dans la mesure du possible. En situation d'incertitude, des politiques figées peuvent créer une vulnérabilité si les hypothèses s'avèrent erronées ou si les conditions changent. Les approches de planification fondée sur des hypothèses et de prise de décision en situation d'incertitude profonde mettent l'accent sur le maintien d'une flexibilité permettant aux politiques publiques d'évoluer. **Les analystes de politiques doivent établir des indicateurs et des seuils de déclenchement signalant la nécessité d'un changement de politique publique, et collecter des données pour suivre les tendances.**

Adapter les cadres de gouvernance de l'IA pour suivre le rythme de l'innovation : le cas de l'IA agentique



Depuis le lancement de grands modèles linguistiques tels que ChatGPT en 2022, le développement d'agents d'IA de plus en plus performants a progressé rapidement, faisant évoluer les systèmes d'IA de la simple réponse aux demandes des utilisateurs vers l'exécution autonome de tâches et l'interaction avec d'autres systèmes. Les agents d'IA représentent un élargissement fondamental de la gouvernance de l'IA, allant au-delà des systèmes individuels déclenchés par les utilisateurs pour inclure les interactions entre plusieurs systèmes autonomes. Comme le Groupe d'experts et d'expertes scientifique international indépendant le souligne : « Le défi en matière de gouvernance passe des modèles d'IA aux agents d'IA. Une défaillance locale pourrait se propager au-delà des frontières organisationnelles et nationales et [...] la sécurité de l'IA pourrait devenir une question de sécurité collective autant que de gouvernance d'entreprise ». Cela nécessite un ajustement des cadres de politiques existants en matière d'IA.

Les défaillances des systèmes agentiques se propagent de manière plus dangereuse que celles des systèmes non agentiques, car les erreurs commises par un seul agent peuvent rapidement se propager en cascade et s'amplifier à travers ces réseaux complexes, souvent sans qu'il soit possible d'intervenir humainement. De tels systèmes sont sujets à de nouveaux modes de défaillance, tels que la mauvaise coordination, où les agents ne parviennent pas à coopérer malgré des objectifs alignés, ou la collusion, où ils se coordonnent d'une manière qui porte atteinte à l'intérêt général. Une expérience a montré que la combinaison de deux modèles « sûrs » pris isolément produisait un logiciel vulnérable dans 43 % des cas, alors que ces mêmes modèles ne présentaient de défaillance que dans moins de 3 % des cas lorsqu'ils fonctionnaient séparément. Il existe également

des risques liés au fait que les agents d'IA mènent des actions de plus en plus autonomes, comme expliqué dans le cas d'utilisation précédent.



Les cadres actuels de gouvernance de l'IA, tels que la loi européenne sur l'IA, ont été adoptés avant l'essor de l'IA agentique et ne sont donc pas adaptés à ces risques. Ils reposent plutôt sur l'hypothèse selon laquelle les fournisseurs individuels peuvent évaluer de manière pertinente les risques relevant de leur propre sphère de contrôle. Dans les environnements multi-agents, cela conduit à une visibilité fragmentée et à des rapports cloisonnés, car aucun acteur ne perçoit comment son système interagit au-delà des frontières organisationnelles. De plus, les tests existants préalables à la mise sur le marché se concentrent sur les propriétés intrinsèques des modèles individuels, ce qui ne permet pas de détecter les dynamiques néfastes qui n'apparaissent que lors des interactions entre prestataires. Enfin, la difficulté d'attribuer la responsabilité en cas d'échec complique la question de la responsabilité; des études montrent que même l'IA de pointe et les experts humains peinent à identifier avec précision quel agent spécifique d'un réseau est à l'origine de l'échec d'une tâche.

L'essor rapide de l'IA agentique montre à quel point les analystes de politiques doivent surveiller et évaluer en permanence les risques liés à l'IA, et adapter leur approche de politique publique à mesure que de nouvelles données apparaissent. Cela plaide également en faveur d'un éventail d'outils de politiques publiques: les cadres juridiques sont difficiles à modifier mais offrent une sécurité aux entreprises; les réglementations peuvent être modifiées plus facilement tout en imposant des exigences obligatoires; enfin, les normes et les lignes directrices sont volontaires et peuvent être ajustées plus rapidement. C'est la voie choisie par Singapour avec le lancement, en janvier 2026, d'un « cadre modèle de gouvernance de l'IA agentique », un document évolutif.

Signalement et enquête sur les incidents de sécurité

Tirer les leçons des incidents liés à la sécurité et des quasi-accidents est essentiel pour réduire l'incertitude et améliorer l'efficacité des politiques au fil du temps. Les mécanismes de signalement des incidents et des quasi-incidents, y compris la protection des lanceurs d'alerte, doivent être associés à des processus d'enquête systématiques afin d'identifier les causes sous-jacentes et les enseignements à en tirer. C'est précisément le principe qui sous-tend les enquêtes sur les accidents aériens et les incidents cybernétiques, et qui constitue la pierre angulaire des [systèmes de gestion de la sécurité](#) et de l'[analyse des causes profondes](#). **Des systèmes de signalement efficaces aident les analystes de politiques à transformer les événements imprévus en opportunités d'apprentissage pour améliorer la sécurité.**

Étayer les affirmations relatives à la sécurité par des preuves

Toute affirmation en matière de sécurité doit s'appuyer sur des preuves transparentes. Ce principe est incarné par les [méthodologies des « dossiers d'assurance »](#) qui exigent des organisations qu'elles construisent des arguments structurés reliant les affirmations de sécurité à des preuves à l'appui telles que des résultats de tests, des analyses formelles, des données opérationnelles ou des normes de certification. Les dossiers d'assurance sont particulièrement utiles lorsque le risque ne peut être entièrement appréhendé de manière probabiliste, car ils permettent de combiner des preuves quantitatives et qualitatives en une justification cohérente. Les dossiers d'assurance constituent le fondement d'un écosystème d'assurance, un réseau d'acteurs, d'outils, de normes et d'institutions conçu pour auditer et valider de manière indépendante les affirmations en matière de sécurité. **En rendant le raisonnement explicite et traçable, les dossiers d'assurance aident les analystes de politiques, les régulateurs et les instituts de sécurité de l'IA à évaluer si les affirmations des entreprises en matière de sécurité de l'IA sont suffisamment étayées malgré l'incertitude et le caractère incomplet des connaissances.**

Communiquer et expliquer les risques

La gestion des risques ne repose pas uniquement sur l'analyse technique, mais aussi sur la manière dont les risques sont interprétés, compris et pris en compte par les parties prenantes. Une communication efficace aide les analystes de politiques, les experts, les parties prenantes et le grand public à développer une compréhension commune de l'incertitude, des dommages potentiels et des réponses possibles. Ce principe se reflète dans le [cadre de l'« amplification sociale du risque »](#) et dans les approches de [« construction du sens du risque »](#),

qui soulignent que les perceptions du risque peuvent influencer de manière significative les comportements et les résultats des politiques. **Une communication claire et appropriée de la part des analystes de politiques peut réduire les malentendus parmi les parties prenantes et le grand public, renforcer la confiance et favoriser une élaboration de politiques plus éclairée en situation d'incertitude.**

Adopter une approche de précaution face à des dommages irréversibles

Lorsque les conséquences d'un risque sont critiques et irréversibles, même une faible probabilité de préjudice peut justifier une action préventive. Le [principe de précaution](#) stipule qu'un manque de certitude scientifique ne doit pas servir de prétexte pour retarder la mise en place de mesures préventives. Il s'applique tout particulièrement lorsque les probabilités sont difficiles à estimer mais que les conséquences potentielles sont catastrophiques ou irréversibles, rendant la prévention préférable à l'attente de preuves concluantes une fois le préjudice survenu. Ce principe est largement appliqué dans la gouvernance environnementale pour protéger les espèces menacées, en santé publique pour éviter des répercussions irréparables sur la santé humaine, et en biotechnologie pour prévenir des dommages durables aux écosystèmes. **Pour les analystes de politiques, si le risque de préjudice est grave et irréversible, il vaut mieux faire preuve de prudence afin d'éviter des conséquences irréparables une fois qu'elles se sont produites, notamment des atteintes graves à la santé humaine, à la vie et à l'environnement.**

Enseignements tirés d'autres domaines de la sécurité et d'autres secteurs de politiques

Au cours des dernières décennies, les analystes de politiques, les régulateurs et les acteurs du secteur ont mis en œuvre divers cadres pour gérer les risques dans des secteurs tels que l'aviation, la sécurité automobile, l'énergie nucléaire, la cybersécurité, le changement climatique, la stabilité financière, la santé publique et la biotechnologie. Bien que ces domaines diffèrent par la nature de leurs dangers, ils sont confrontés à bon nombre des mêmes défis qui caractérisent aujourd'hui les risques liés à l'IA : probabilités incertaines, interactions complexes entre les systèmes, technologies en constante évolution et possibilité d'événements à faible probabilité mais aux conséquences graves. L'examen de la manière dont ces secteurs ont mis en œuvre les principes présentés dans la section précédente peut fournir des enseignements utiles pour la gestion des risques liés à l'IA en situation d'incertitude.

Sécurité aérienne

Dans le domaine de l'aviation, [la gestion des risques s'appuie sur des systèmes de gestion de la sécurité](#) et des méthodes d'évaluation probabiliste des risques qui visent à identifier, surveiller et atténuer les dangers avant qu'ils ne se traduisent par des accidents. Plutôt que de partir du principe que les défaillances peuvent être entièrement éliminées, les autorités aéronautiques et les exploitants analysent la probabilité et les conséquences potentielles des dysfonctionnements techniques, des erreurs humaines, des phénomènes météorologiques et des défaillances organisationnelles dans un large éventail de scénarios opérationnels. La sécurité aérienne moderne s'appuie également fortement sur la collecte continue de données, notamment les rapports d'incidents, la surveillance des données de vol et l'analyse des quasi-accidents, afin de détecter les signaux faibles et les risques émergents au sein de systèmes complexes. Cela permet aux régulateurs et aux compagnies aériennes d'adapter leurs politiques de sécurité, de donner la priorité aux interventions là où les risques sont les plus élevés et de renforcer la résilience du système face à des événements rares mais catastrophiques, tels que les défaillances en cascade d'équipements ou les accidents par perte de contrôle. **La sécurité aérienne peut enseigner aux analystes de politiques en matière d'IA l'importance d'identifier et de prévenir les causes potentielles de dangers, ainsi que de tirer les leçons des incidents pour renforcer les mesures préventives.**

Sécurité automobile

Dans l'industrie automobile, [l'approche dominante en matière de gestion des risques est la sécurité fonctionnelle](#), qui combine des normes de sécurité techniques, des essais de collision et une analyse des risques basée sur les systèmes afin de réduire à la fois la probabilité d'accidents et la gravité des blessures. Les constructeurs ont recours à des méthodes d'essais déterministes, telles que les essais de collision standardisés (frontaux, latéraux et de renverse-

ment), parallèlement à une analyse statistique des données d'accidents réels et des modes de défaillance. Plutôt que de partir du principe que les collisions peuvent être entièrement évitées, la sécurité des véhicules s'articule autour d'un « système sûr » à plusieurs niveaux qui intègre des dispositifs de protection tels que les ceintures de sécurité, les coussins gonflables, les systèmes de freinage antiblocage et des systèmes d'aide à la conduite de plus en plus sophistiqués. La sécurité automobile moderne s'appuie également sur des approches probabilistes et d'ingénierie des systèmes, en particulier pour les véhicules autonomes à forte intensité logicielle, où les comportements humains, les erreurs des capteurs et les interactions complexes entre les conducteurs et les systèmes automatisés doivent être évalués dans un large éventail de scénarios. Ensemble, ces méthodes d'essai, ces couches de protection et ces analyses fondées sur les données réduisent les risques d'accident et de blessure dans diverses conditions de conduite. **Cela souligne l'importance pour les analystes de politiques d'anticiper les scénarios de dangers potentiels avant le déploiement des modèles et d'exiger des mesures visant à atténuer leur impact.**

Sûreté nucléaire

Le risque lié à la sûreté nucléaire est principalement abordé par le biais de l'évaluation probabiliste des risques qui modélise les interactions entre les défaillances techniques, les erreurs humaines et les aléas externes (par exemple les tremblements de terre ou les inondations) afin d'estimer la probabilité d'accidents graves. Cette approche prend en compte les événements à faible probabilité mais aux conséquences importantes, y compris les défaillances en cascade pouvant entraîner des rejets radioactifs. La sûreté nucléaire recourt également à l'analyse déterministe de sûreté, dans le cadre de laquelle les ingénieurs évaluent si un réacteur peut résister à des « accidents de référence » prédéfinis en utilisant des hypothèses prudentes. Contrairement à l'analyse probabiliste des risques, l'analyse déterministe de sûreté ne cherche pas à déterminer la probabilité qu'un événement se produise, mais à vérifier si la centrale peut y résister en toute sécurité s'il venait à se produire. C'est pourquoi les réacteurs nucléaires intègrent de multiples barrières indépendantes et redondantes pour empêcher tout rejet radioactif. Ce principe de « défense en profondeur » repose sur l'idée qu'aucune défaillance isolée ne doit conduire directement à une catastrophe. En analysant un large éventail de scénarios possibles, les ingénieurs peuvent hiérarchiser les investissements en matière de sûreté et renforcer la redondance ainsi que la préparation aux situations d'urgence.

Cela rappelle aux analystes de politiques du domaine de l'IA qu'ils doivent évaluer le risque d'effets en cascade et planifier des lignes de défense à plusieurs niveaux afin d'éviter des conséquences catastrophiques.

Cybersécurité

La cybersécurité repose sur une approche de gestion continue des risques qui vise à identifier, évaluer, atténuer et surveiller les menaces pesant sur les

systèmes informatiques face à des adversaires en constante évolution. Plutôt que de partir du principe que les attaques peuvent être totalement évitées, la cybersécurité met l'accent sur la résilience grâce à des défenses en plusieurs couches, une surveillance continue, le renseignement sur les menaces, la réponse aux incidents et une reprise rapide. Les évaluations des risques prennent généralement en compte la probabilité que des acteurs malveillants spécifiques exploitent des vulnérabilités connues, les conséquences potentielles d'attaques réussies et l'efficacité des contrôles de sécurité existants. Des techniques, telles que la modélisation des menaces et des primes aux bogues qui font appel à des hackers éthiques, aident les organisations à anticiper les voies d'attaque potentielles, à corriger les failles de sécurité et à hiérarchiser les mesures défensives dans des environnements technologiques en évolution rapide où de nouvelles menaces et vulnérabilités apparaissent sans cesse. **Face à un environnement de l'IA qui évolue tout aussi rapidement, les analystes de politiques devraient surveiller en permanence les menaces potentielles et prévoir une réponse rapide en cas d'incident.**

COVID-19

Pendant la pandémie de COVID-19, les analystes de politiques ont dû faire face à une situation marquée par une profonde incertitude, dans laquelle des facteurs critiques, tels que les taux de transmission, la dynamique des mutations, l'efficacité des vaccins et les comportements du public, ne pouvaient être prédits de manière fiable à l'avance. Les gouvernements ont dû passer d'une approche « prédire et agir » à une approche « se préparer, surveiller et s'adapter ». Ils ont identifié des indicateurs de résultats préétablis, tels que le taux officiel de positivité des tests, l'activité virale dans les eaux usées, les admissions à l'hôpital liées à la COVID-19, le taux d'occupation des lits d'hospitalisation et les décès liés à la COVID-19, auxquels ont été associées des mesures d'urgence spécifiques pour gérer les risques sanitaires, telles que le port obligatoire du masque, le dépistage et la recherche des contacts, la limitation des rassemblements, l'augmentation de la capacité des établissements de santé, l'annulation des interventions non urgentes, etc. En surveillant des indicateurs spécifiques qui déclenchaient des mesures d'urgence plutôt qu'en optimisant les mesures sanitaires sur la base de conditions prévues, les gouvernements ont amélioré la robustesse de leur réponse face à toute une gamme de scénarios possibles. Cette approche, appelée « prise de décision en situation d'incertitude profonde », aide les analystes de politiques à surmonter le « dilemme de la preuves », c'est-à-dire le défi consistant à décider s'il faut agir sur la base d'informations limitées ou prendre le risque d'attendre des données concluantes qui pourraient arriver trop tard pour prévenir des dommages. Elle a permis aux institutions de prendre des décisions proactives et éclairées dans un environnement incertain où les données restaient insuffisantes pour caractériser pleinement les conditions futures. **L'approche adoptée pour répondre à la pandémie de COVID-19 devrait encourager les analystes de politiques en matière d'IA à surveiller des indicateurs spécifiques, tels que le taux de chômage, qui déclenchent des mesures d'urgence, comme l'ajustement des programmes de transition professionnelle.**

Risque financier

À la suite de la crise financière de 2008-2009, [les analystes de politiques ont commencé à s'appuyer sur la gestion des risques systémiques](#) pour identifier et atténuer les risques dans le secteur financier. Cette approche reconnaît que les risques au niveau des entreprises et les chocs locaux peuvent potentiellement entraîner des perturbations du système, une instabilité généralisée et une crise financière, car ils peuvent se propager à travers un réseau d'institutions et de marchés interconnectés. Les régulateurs et les banques centrales utilisent une combinaison d'outils macroprudentiels, tels que les tests de résistance et la surveillance des réseaux, pour évaluer les vulnérabilités à l'échelle du système financier. L'objectif est d'évaluer comment plusieurs institutions se comporteraient face à des chocs communs, plutôt que de manière isolée. La gestion des risques systémiques fournit une base structurée pour concevoir des interventions, telles que des réserves de fonds propres, des exigences de liquidité et des mécanismes de prêt d'urgence, qui renforcent la résilience financière, réduisent le risque de contagion et garantissent que les fonctions financières essentielles puissent continuer à fonctionner même en période de tensions marquées sur les marchés. **De même, les analystes de politiques ne devraient pas seulement évaluer les risques liés aux modèles d'IA de manière isolée, mais également les vulnérabilités découlant des interactions entre ces modèles, en particulier avec l'essor des agents d'IA, et planifier des mesures préventives et d'atténuation en conséquence. Cela s'applique également aux risques systémiques mentionnés ci-dessus.**

Changement climatique

Le Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC) tient compte de l'incertitude liée à l'évaluation des conséquences du changement climatique en [recourant à toute une série de scénarios futurs](#). Ces scénarios, fondés sur différentes hypothèses concernant la croissance démographique, le développement technologique, la consommation d'énergie et les émissions de gaz à effet de serre, permettent aux scientifiques de modéliser l'évolution possible du climat en fonction de différents choix de politiques. Le GIEC étudie également les « points de basculement » potentiels, tels que l'effondrement irréversible des calottes glaciaires ou la perturbation à grande échelle des écosystèmes, où un réchauffement progressif pourrait déclencher des changements soudains et potentiellement catastrophiques. En combinant l'analyse de scénarios et l'évaluation des risques, le GIEC fournit aux analystes de politiques un cadre permettant d'évaluer les coûts et les bénéfices des stratégies d'atténuation et d'adaptation en situation d'incertitude. Cela aide les gouvernements à concevoir des politiques qui restent solides face à de multiples avenir possibles et qui réduisent la probabilité de franchir des seuils climatiques dangereux. Cependant, malgré des efforts considérables pour réaliser des modélisations climatiques sophistiquées basées sur des données historiques, l'incertitude concernant les trajectoires du changement climatique reste importante; des écarts par rapport aux modélisations sont possibles. **Indépendamment de cette limite, l'approche du GIEC peut apprendre aux analystes spécialisés en IA à concevoir des politiques qui restent solides face à un large éventail de scénarios.**

Sécurité des médicaments

Avant qu'un nouveau médicament puisse être commercialisé à grande échelle, les autorités de régulation exigent généralement que des données soient recueillies progressivement, en commençant par des études précliniques, puis en passant par des phases d'essais cliniques impliquant des populations de plus en plus importantes. Les premiers essais portent principalement sur la sécurité et la posologie, tandis que les phases ultérieures évaluent l'efficacité, les effets indésirables et le rapport entre bénéfices et risques. Les autorités de régulation ont recours à des techniques telles que les essais par étapes, les seuils de sécurité prédéfinis et les règles d'arrêt, les essais contrôlés, l'examen indépendant et les restrictions concernant les personnes autorisées à participer ou à recevoir le médicament, afin de limiter l'exposition tant que les données sont encore en cours de collecte. L'autorisation de mise sur le marché ne met pas fin à ce processus : les autorités de régulation peuvent exiger une surveillance post-commercialisation, la déclaration des événements indésirables, des études supplémentaires, des plans de gestion des risques, des modifications de la notice ou des restrictions d'utilisation à mesure que de nouvelles données apparaissent. **Il s'agit là d'un autre exemple de gestion adaptative des risques destiné aux analystes de politiques en matière d'IA : l'exposition est progressivement étendue à mesure que les données s'accumulent, tandis que la surveillance et les mesures de protection réglementaires après la mise en marché permettent de détecter les effets néfastes et d'ajuster ou de retirer l'autorisation si nécessaire.**

Organismes génétiquement modifiés (OGM)

La gestion des risques liés aux organismes génétiquement modifiés (OGM) repose sur une combinaison d'évaluation scientifique des risques et d'évaluation fondée sur des données probantes, de réglementation préventive et de surveillance, afin d'évaluer les impacts potentiels sur la santé humaine, l'agriculture et l'environnement. Les autorités de régulation procèdent généralement à des évaluations au cas par cas qui examinent des questions telles que la toxicité, l'allergénicité, le transfert de gènes, les effets écologiques et les impacts sur la biodiversité avant d'approuver la culture ou la consommation d'OGM. Étant donné que certains risques peuvent ne pas être pleinement compris avant la mise en œuvre, notamment en ce qui concerne les interactions écologiques à long terme, de nombreux cadres réglementaires intègrent également des éléments du principe de précaution, exigeant des tests supplémentaires, des mesures de confinement ou une mise en œuvre progressive lorsque les données sont incomplètes. Une surveillance continue et un suivi environnemental sont souvent mis en œuvre pour détecter les effets indésirables et actualiser les évaluations des risques à mesure que de nouvelles informations deviennent disponibles. **La gestion des risques liés aux OGM offre un exemple concret aux analystes de politiques en matière d'IA que, face à des données incomplètes et à un risque de conséquences irréversibles, une approche de précaution peut être adoptée.**

L'examen des approches de gestion des risques et de prise de décision révèle qu'il n'existe pas de méthode unique capable de traiter la gestion des risques en situation d'incertitude. Une gouvernance efficace repose plutôt sur une combinaison d'approches complémentaires, notamment des défenses en plusieurs niveaux, une planification adaptative, un suivi continu, l'analyse de scénarios et des mesures de précaution, comme le préconise également le Rapport international sur la sécurité de l'IA. Alors que la gestion des risques nous indique comment gérer les dangers, les analystes de politiques ont besoin d'un cadre pour déterminer quand, pourquoi et comment les gouvernements doivent intervenir.

Segmenter : L'élaboration de politiques en matière d'IA en situation d'incertitude

Les systèmes d'IA à usage général s'inscrivent au sein de systèmes socio-techniques complexes dans lesquels les risques peuvent prendre naissance à différents niveaux, se propager à travers ces niveaux et interagir de manière inattendue. Par conséquent, une élaboration efficace des politiques nécessite des décisions à plusieurs niveaux du système. L'objectif n'est pas d'éliminer tous les risques, mais d'accroître la résilience de l'ensemble du système en réduisant les vulnérabilités, en limitant la propagation des préjudices et en préservant la capacité d'adaptation à mesure que les connaissances et les conditions évoluent.

Qu'entend-on par « résilience » ?

- **La résistance** réduit la probabilité ou la gravité d'un choc avant qu'il ne se produise.
- **L'absorption** permet aux systèmes sociétaux de maintenir leurs fonctions essentielles pendant un choc.
- **La reprise** aide à rétablir le fonctionnement normal après la survenue d'un choc.
- **L'adaptation** transforme les systèmes sociétaux afin de réduire leur vulnérabilité face à de futurs chocs.

Source : Bengio, Y. et al. (2026). *Rapport international sur la sécurité de l'IA*.

S'appuyant sur l'idée que les analystes de politiques peuvent influencer des systèmes complexes par des interventions à plusieurs niveaux, comme évoqué dans « [The Forever Crisis : Adaptive Global Governance for an Era of Accelerating Complexity](#) », cette boîte à outils fournit un cadre permettant d'identifier et de combiner des interventions publiques à différents niveaux. Elle distingue quatre niveaux auxquels des risques liés à l'IA peuvent émerger : le niveau des modèles, le niveau individuel, le niveau sociétal et le niveau international.

Identifier le niveau auquel un risque trouve principalement son origine peut aider les analystes de politiques à déterminer où une intervention a le plus de chances d'être efficace et efficiente. Il est important de noter que ces niveaux sont interconnectés : les faiblesses à un niveau peuvent se propager aux autres. Par conséquent, la résilience ne dépend pas d'une mesure de politique isolée, mais de l'efficacité combinée des interventions à tous les niveaux, comme le soulignent le [Rapport international sur la sécurité de l'IA](#) et [le rapport préliminaire du Panel scientifique international indépendant sur l'IA](#).



Niveau du modèle :

Sécurité technique, fiabilité et alignement



Niveau de l'individu :

Prévention des utilisations abusives



Niveau sociétal :

Résilience systémique



Niveau international :

Coordination

** Ce cadre identifie le niveau auquel les risques peuvent apparaître ainsi que l'objectif stratégique associé aux interventions de politiques à ce niveau.*



Niveau du modèle : Sécurité technique, fiabilité et alignement

Source des risques :

Au niveau du modèle, les risques proviennent de défaillances techniques, de vulnérabilités, des décisions des développeurs, de données biaisées, de comportements inattendus au sein même des modèles et de défaillances techniques entre les agents d'IA. Des risques peuvent également découler de la possibilité que de futurs modèles d'IA à usage général présentent des capacités, des comportements ou des interactions fondamentalement difficiles à prévoir et susceptibles d'entraîner un décalage entre les comportements des systèmes d'IA et les valeurs humaines. Cela pourrait inclure des scénarios de perte de contrôle et d'autres issues inconnues pouvant avoir des conséquences irréversibles.

Type d'incertitude :

Les analystes de politiques sont confrontés à deux formes distinctes d'incertitude à ce niveau. La première est l'incertitude aléatoire, qui reflète la variabilité statistique inhérente aux modèles d'IA. La seconde est l'incertitude émergente, qui découle des progrès et des interactions imprévisibles des capacités de l'IA. Une grande partie des données nécessaires à la compréhension de la conception et des capacités des modèles d'IA à usage général est générée et détenue par des entreprises. Pour réduire l'incertitude, les analystes de politiques et les régulateurs doivent donc avoir accès à [des informations standardisées provenant de l'ensemble des entreprises](#) afin de pouvoir évaluer de manière cohérente la conception et les capacités des modèles, analyser les pratiques de sécurité, identifier les risques émergents et déterminer si les mesures de gouvernance existantes restent adéquates.

Prévention et atténuation :

Les risques au niveau des modèles peuvent être atténués grâce à des choix de conception, des méthodes d'ingénierie et divers niveaux de défense. Les enseignements tirés de l'aviation, de la sécurité automobile, de la cybersécurité et de l'énergie nucléaire suggèrent que les analystes de politiques devraient se concentrer sur la mise en place de pratiques de « sécurité dès la conception ». [L'analyse des modes de défaillance et de leurs effets, les cas d'assurance, l'évaluation déterministe de la sécurité, la défense en profondeur et l'analyse des causes profondes](#) partagent toutes un postulat commun : des défaillances se produiront et, par conséquent, les systèmes doivent être conçus pour les détecter, les contenir et s'en remettre avant qu'elles n'entraînent des dommages importants. Face à des dommages irréversibles, les analystes de politiques doivent s'appuyer sur une [approche de précaution](#). Cette approche met l'accent sur la prudence face à une forte incertitude scientifique, à des conséquences critiques et à des dommages potentiellement irréversibles, ce qui peut justifier une intervention publique avant même que des preuves n'apparaissent. Dans de nombreux domaines où la sécurité est cruciale, notamment l'aviation, la sûreté nucléaire et la préparation aux pandémies, les décideurs de politiques n'ont pas attendu que des incidents se produisent pour mettre en place des mesures de protection face au risque de dommages irréversibles tels que la perte de vies humaines. Au contraire, ils ont agi pour réduire la probabilité et les conséquences d'incidents potentiellement critiques, reconnaissant que le coût de l'inaction pourrait largement dépasser celui d'une intervention préventive.

Objectif des politiques :

L'objectif à ce niveau est de renforcer la sécurité technique, la fiabilité et la conformité des systèmes d'IA en favorisant une gestion efficace des risques, la transparence et la responsabilité, le développement et le déploiement responsables des capacités, ainsi que la préparation aux interventions d'urgence. Cela peut se faire par le biais de recommandations volontaires et d'exigences obligatoires.

Exemples d'intervention :

Publication de fiches de modèles, signalement et enquête sur les incidents et les quasi-incidents, évaluation et analyse comparative normalisées de la sécurité, normes techniques, exercices de simulation d'attaques (red-teaming), programme de prime aux bogues, protocoles de sécurité « If-Then » et contrôles au niveau du système, évaluation ou audit indépendant, protection pour les lanceurs d'alerte, bacs à sable réglementaires, contrôles d'urgence et financement de la recherche.



Niveau individuel : Prévention des utilisations abusives

Source des risques :

Au niveau individuel, les risques découlent de la manière dont les utilisateurs, les organisations ou les acteurs malveillants déploient, exploitent et interagissent avec les systèmes d'IA. Les vulnérabilités des systèmes d'IA peuvent être exploitées pour altérer le modèle et supprimer les mesures de protection, faciliter les cyberattaques, la fraude, les opérations d'influence, les activités criminelles ou le développement de menaces biologiques et chimiques. Les analystes de politiques sont souvent confrontés à des informations incomplètes concernant les nouvelles voies d'utilisation abusive, les comportements des utilisateurs, ainsi que les capacités et les modes d'utilisation des attaquants, ce qui rend l'élaboration des politiques plus difficile.

Type d'incertitude :

À ce niveau, les analystes de politiques sont principalement confrontés à une incertitude épistémique, résultant d'une connaissance incomplète des nouvelles voies d'utilisation abusive, des comportements des utilisateurs, ainsi que des capacités et des modes d'utilisation des attaquants. Cette incertitude peut être réduite grâce à la recherche, au signalement des incidents, au partage d'informations, aux environnements de test (sandboxes) et à la surveillance, ce qui permet aux analystes de politiques de recueillir davantage de données sur les voies d'utilisation abusive et leurs conséquences.

Prévention et atténuation :

Ces risques impliquant souvent des adversaires agissant de manière intentionnelle, les approches empruntées à la cybersécurité sont particulièrement pertinentes. La [modélisation des menaces](#), [l'analyse Bowtie](#) et [l'analyse « What-if »](#) aident à identifier les voies d'attaque, à anticiper les scénarios d'utilisation abusive et à mettre en place des contrôles préventifs et d'atténuation. Le signalement et l'enquête sur les incidents dans le cadre du [système de gestion de la sécurité](#), ainsi que [l'analyse des causes profondes](#) utilisée en sécurité aérienne, peuvent également contribuer à recueillir des preuves et à comprendre les causes sous-jacentes. Plutôt que de se concentrer sur les défaillances techniques, ces approches examinent comment les systèmes peuvent être délibérément exploités et comment les mesures de protection peuvent être renforcées en conséquence.

Objectif des politiques :

L'objectif à ce niveau est de réduire les possibilités d'utilisation abusive grâce à la transparence, à la gestion des risques et à la mise en place de mesures de protection.

Exemples d'intervention :

Littératie de l'IA, l'étiquetage et le filigrane numérique des contenus générés par l'IA, mécanisme d'authentification des contenus, cybersécurité et garde-fous contre les utilisations abusives, exigences en matière de connaissance du client, systèmes de surveillance de l'utilisation et de détection des abus, et partage des failles de sécurité et des renseignements sur les menaces avec d'autres acteurs du secteur. L'application des cadres juridiques existants peut également venir compléter ces outils.



Niveau sociétal : Résilience systémique

Source des risques :

Au niveau sociétal, les risques découlent des effets cumulés de l'adoption de l'IA par les individus au sein des organisations, des marchés et des systèmes sociaux. Ces préjudices ne sont pas nécessairement imputables à un modèle ou à un utilisateur en particulier, mais résultent plutôt des comportements des acteurs agissant au sein de systèmes interconnectés. On peut citer, à titre d'exemples, la perturbation du marché du travail, l'érosion de l'autonomie humaine et les répercussions sur les processus démocratiques et les écosystèmes de l'information. Les analystes de politiques manquent souvent de données fiables sur le taux d'adoption ou les impacts à long terme. Ils peuvent également s'appuyer sur des hypothèses erronées ou incomplètes concernant la manière dont la société s'adaptera, ou rencontrer des difficultés à établir le lien de causalité entre les capacités de l'IA et les préjudices sociétaux.

Type d'incertitude :

Les principales incertitudes à ce niveau sont l'incertitude épistémique, l'incertitude liée aux modèles et l'ambiguïté. Les analystes de politiques manquent souvent de données fiables sur le taux d'adoption ou les impacts à long terme (incertitude épistémique). Ils peuvent s'appuyer sur des hypothèses erronées ou incomplètes concernant le fonctionnement du système, ou bien les relations de causalité entre les capacités de l'IA et les préjudices sociétaux peuvent être difficiles à établir (incertitude liée aux modèles). Il peut également ne pas y avoir de consensus au sein des cercles de politiques et de recherche quant à la définition, l'interprétation ou la formulation des concepts clés (ambiguïté). Pour réduire ces incertitudes, les analystes de politiques devraient mettre au point des indicateurs d'alerte précoce permettant de suivre les tendances sociétales, telles que les bouleversements du marché du travail liés à l'adoption de l'IA, et financer des recherches axées sur la relation de causalité entre les capacités de l'IA et les préjudices. Ils devraient également collaborer avec des chercheurs pour définir des concepts clés, tels que l'autonomie humaine, réduisant ainsi l'ambiguïté.

Prévention et atténuation :

Des approches telles que la gestion des risques systémiques, l'analyse de l'effet domino, la planification fondée sur des hypothèses, les trajectoires de politiques adaptatives dynamiques, l'amplification sociale du risque et la construction du sens du risque s'avèrent donc particulièrement utiles. Ces méthodes encouragent les analystes de politiques à analyser les interdépendances, à identifier les hypothèses critiques, à surveiller les tendances émergentes en identifiant des indicateurs et des points de déclenchement et à préparer des réponses adaptatives à mesure que les conditions sociétales évoluent.

Objectif des politiques :

L'objectif à ce niveau n'est pas de prévenir les préjudices individuels, mais de renforcer la résilience sociétale et la protection des droits.

Exemples d'intervention :

Programmes de littératie numérique et en IA, réformes des lois sur la protection des consommateurs et la vie privée, règles relatives aux contenus générés par l'IA dans un contexte électoral, programmes de formation continue et de perfectionnement, programmes d'aide à la transition professionnelle, indicateurs relatifs à l'IA, observatoires de l'IA et centres de recherche.



Niveau international : Coordination

Source des risques :

Au niveau international, les risques découlent de la nature transnationale des conséquences des risques liés à l'IA : la fraude et les cyberattaques ne respectent pas les frontières nationales ; une attaque biochimique rendue possible par l'IA pourrait être mise au point dans un pays et déployée dans un autre ; et du matériel pédopornographique ou des images intimes non consenties peuvent être générés dans une juridiction et diffusés à l'échelle mondiale. Malgré ces risques transfrontaliers, il n'existe pas de définitions convenues au niveau international pour des concepts tels que l'IA de pointe, les seuils de capacité, les indicateurs d'évaluation de la sécurité ou le risque catastrophique. Les politiques nationales élaborées de manière isolée peuvent en réalité créer des exigences incohérentes et entraîner des obligations de conformité redondantes ou qui se chevauchent pour les entreprises. Par exemple, un algorithme de recrutement considéré comme légal à Singapour peut nécessiter une évaluation de conformité au titre de la loi européenne sur l'IA et un audit algorithmique au titre de la loi locale 144 de la ville de New York.

Type d'incertitude :

Les principales formes d'incertitude à ce niveau sont l'incertitude liée aux modèles et l'ambiguïté. Bien que certaines instances intergouvernementales telles que l'ASEAN, le Conseil de l'Europe, l'UE, l'Union africaine, l'OEA, l'OCDE et l'ONU discutent de la gouvernance de l'IA, il reste difficile pour les gouvernements de savoir comment d'autres juridictions réglementeront des systèmes d'IA aux capacités croissantes et comment la concurrence géopolitique pourrait influencer les choix réglementaires (incertitude liée aux modèles). Parallèlement, l'ambiguïté découle de l'absence de définitions et de cadres de mesure convenus au niveau international pour des concepts tels que l'IA de pointe, les seuils de capacité, la sécurité ou le risque catastrophique. Ces incertitudes peuvent être réduites grâce à des dialogues et des accords internationaux sur des concepts clés en matière de sécurité, tels que les indicateurs d'évaluation de la sécurité.

Prévention et atténuation :

Les approches issues de la gouvernance internationale en matière de sécurité fournissent des orientations précieuses pour gérer ces défis. Les domaines de [l'aviation et de la sûreté nucléaire démontrent](#) que les technologies à haut risque peuvent être régies au-delà des frontières grâce à des normes minimales communes de sécurité élaborées par l'industrie en partenariat avec les institutions de sûreté, à des audits et au partage d'informations, à la déclaration et à l'enquête mondiales et anonymisées sur les incidents, ainsi qu'à des comparaisons entre pairs visant à promouvoir le respect des normes de sécurité. Ces cadres de gouvernance internationale s'appuient sur les autorités nationales pour faire respecter les règles convenues. Ils soulignent que la coordination internationale n'a pas pour but d'imposer des régimes réglementaires identiques, mais d'établir une base commune en matière de sécurité, d'améliorer l'efficacité de la gouvernance nationale de l'IA et de limiter les contraintes de conformité pour l'industrie. Cependant, l'élaboration de tels cadres prend du temps, ce qui peut limiter leur utilité lorsque les risques pourraient se concrétiser avant qu'un consensus ne soit atteint.

Objectif des politiques :

L'objectif est donc de renforcer la coordination internationale et l'interopérabilité normative.

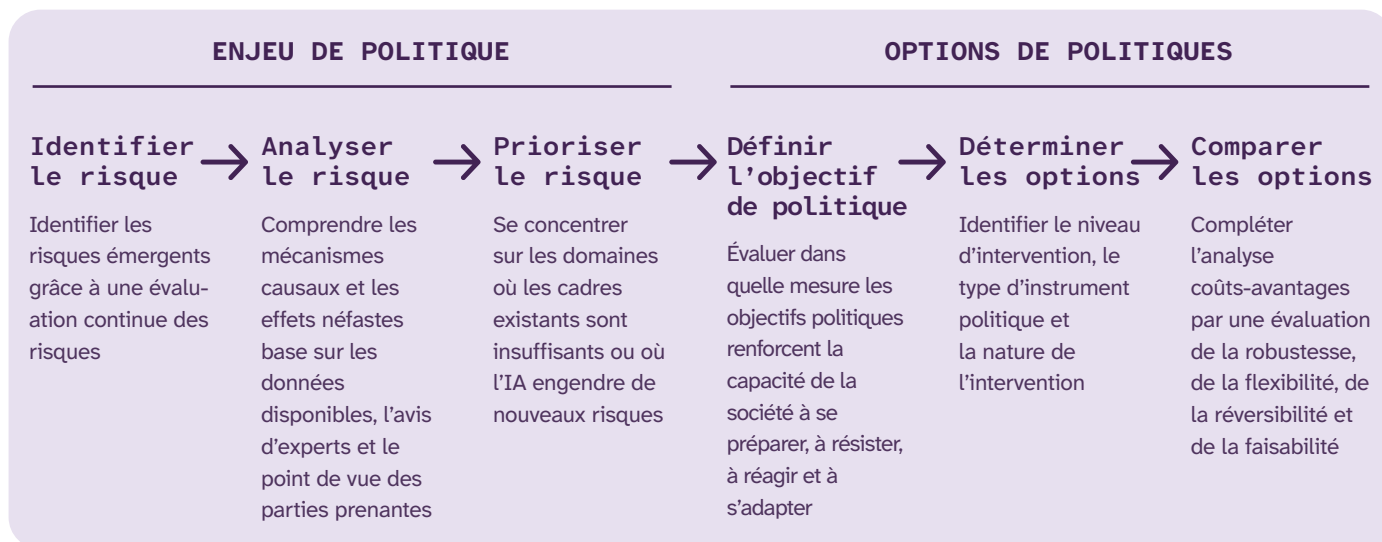
Exemples of intervention :

Groupes d'experts scientifiques internationaux, réseaux internationaux d'instituts de sécurité, systèmes mondiaux de signalement des incidents, rapports internationaux sur la transparence, normes internationales et traités internationaux.

Agir : L'élaboration de politiques en matière d'IA dans la pratique

L'incertitude entourant la probabilité, le moment d'apparition et la gravité de nombreux préjudices liés à l'IA limite la capacité des analystes de politiques à mener une évaluation probabiliste des risques et à déterminer les coûts et les bénéfices d'une intervention publique. Les enseignements tirés de l'analyse, de la gestion et de la prise de décision en matière de risques, ainsi que d'autres domaines critiques pour la sécurité peuvent aider à ajuster le processus par lequel les analystes conçoivent leurs politiques. La section suivante fournit des orientations aux analystes de politiques sur la manière d'intégrer ces enseignements dans l'élaboration des politiques; un résumé de deux pages décrivant les étapes est disponible à l'annexe 4. Il convient de noter que la section suivante ne traite que des étapes analytiques de l'élaboration des politiques; elle n'a pas pour objectif de proposer une voie normative précise aux analystes de politiques qui devront gérer ces choix en fonction de leurs propres contextes et priorités.

Enjeu de politiques : identifier, analyser et hiérarchiser les risques



Identification des risques

La justification d'une intervention commence par l'identification et la description des risques de préjudice liés à l'IA. Les analystes politiques peuvent s'appuyer sur des données scientifiques et des évaluations d'experts, telles que le [Rapport international sur la sécurité de l'IA](#), le [Panel scientifique international indépendant sur l'IA](#), [l'Indice Stanford sur l'IA](#), [l'Initiative du MIT sur les risques liés à l'IA](#) et le Forecasting Research Institute. Cependant, les technologies d'IA évoluent rapidement, tout comme les risques qui y sont associés. Une identification efficace des risques doit donc être un processus continu afin de cerner les risques émergents à mesure que de nouvelles données sont mises à disposition par la recherche universitaire, les reportages journalistiques, les rapports d'incidents et les échanges réguliers avec des experts techniques et socio-techniques issus du monde universitaire, du secteur privé et de la société civile. Les analystes de politiques doivent également expliciter les limites de ces évaluations : compte tenu de l'incertitude qui entoure les risques liés à l'IA, toute évaluation est susceptible d'être incomplète et doit être ré-examinée à mesure que les données et les connaissances évoluent.

Exemple d'identification des risques

Risque de cyberattaque utilisant l'IA contre des infrastructures critiques (par exemple, hôpitaux, ports, aéroports, réseau électrique, approvisionnement en eau)

Analyse des risques

Une fois les risques identifiés, les analystes de politiques devraient s'attacher à comprendre comment ceux-ci pourraient se concrétiser, ce qui peut nécessiter d'examiner les capacités et les vulnérabilités des modèles à un niveau micro afin de déterminer les causes potentielles de défaillances et d'écarts par rapport à la conception prévue du modèle, ainsi que d'analyser à un niveau macro la manière dont les systèmes d'IA sont déployés, utilisés (et détournés) dans des contextes concrets afin de comprendre la chaîne causale menant aux impacts sociétaux en aval. Les rapports d'incidents liés à l'IA peuvent servir de source de données factuelles tant pour les incidents à l'échelle micro que macro. Les analystes de politiques devraient également s'efforcer de comprendre les voies causales par lesquelles les risques liés à l'IA pourraient se traduire par des préjudices, afin d'identifier le point d'intervention et d'atténuer les conséquences potentielles.

Les analystes politiques évaluent ensuite les conséquences potentielles associées à chaque risque identifié. Caractériser la gravité et la réversibilité des conséquences peut aider les analystes de politiques à comprendre l'impact potentiel sur la société. Si l'évaluation de l'ampleur de l'impact de certains

risques, tels que les pertes financières résultant de fraudes ou de cyberattaques basées sur l'IA, peut s'appuyer en partie sur des données existantes, de nombreux risques liés à l'IA ne disposent pas de données empiriques suffisantes pour étayer des estimations quantitatives fiables. Dans ces situations, les évaluations qualitatives, l'avis d'experts et la concertation avec les parties prenantes concernées constituent une base importante pour évaluer les préjudices potentiels.

Il est important de noter que, face à l'incertitude, les analystes de politiques doivent distinguer les hypothèses solides de celles qui sont incertaines qui sous-tendent leur compréhension des causes et des conséquences des risques. Par exemple, des scénarios de « perte de contrôle » pourraient résulter d'interactions entre certaines capacités avancées spécifiques d'un système d'IA. Cependant, les liens de causalité par lesquels ces capacités pourraient conduire à une perte de contrôle restent des hypothèses qui doivent être vérifiées et pourraient devoir être révisées à mesure que les connaissances évoluent. Les mesures politiques destinées à atténuer ces causes sous-jacentes doivent donc tenir compte de l'incertitude de ces hypothèses. Les analystes de politiques peuvent mettre en place des mécanismes permettant de réexaminer et d'ajuster ces mesures si des éléments de preuve démontrent que les hypothèses sur lesquelles elles reposent sont erronées.

Exemple d'analyse des risques

Le risque de cyberattaque facilitée par l'IA peut se manifester tant au niveau du modèle (par exemple, de manière non intentionnelle) qu'au niveau individuel (de manière intentionnelle). Les voies non intentionnelles peuvent inclure des défaillances dans les systèmes d'IA dotés d'autonomie, telles qu'un décalage entre l'intention de l'utilisateur et les comportements du modèle, ou des interactions inattendues entre plusieurs agents d'IA. Les voies intentionnelles nécessitent d'évaluer les acteurs potentiellement menaçants (par exemple, les États-nations, les organisations cybercriminelles, les groupes terroristes) ainsi que leurs capacités, leurs motivations, leurs tactiques et leurs cibles probables. Le préjudice causé par une cyberattaque facilitée par l'IA peut être qualifié de critique mais réversible, car il implique des dommages temporaires ou réparables aux infrastructures sociétales critiques, à moins que la cyberattaque n'entraîne des pertes humaines, auquel cas le préjudice devient irréversible. L'ampleur de l'impact potentiel peut être estimée à partir des données issues de cyberattaques antérieures contre des infrastructures critiques, tout en reconnaissant que les attaques générées par l'IA peuvent différer en termes d'échelle, de vitesse, de sophistication ou de durée.

Tableau 2. Matrice illustrative de classification des risques de préjudice liés à l'IA³

GRAVITÉ DU PRÉJUDICE		RÉVERSIBILITÉ DU PRÉJUDICE	
		Oui	Non
Modérée	Préjudice causé aux individus ou à la société, ayant un impact négatif sur leur capacité à fonctionner ou sur leur qualité de vie, tel que la perte de l'esprit critique, la polarisation, la perte d'emploi et les inégalités économiques	Le préjudice est temporaire ou peut être réparé	Le préjudice est permanent ou irréparable
		Influence et manipulation	
		Impacts sur le marché du travail	
Importante	Préjudice causé à des individus, des groupes ou des organisations, notamment des pertes financières, une atteinte à la réputation, de la discrimination ou des dommages physiques et psychologiques	Risques pour l'autonomie humaine	
		Contenu généré par l'IA et criminalité	
Critique	Préjudice causé à des individus ou à la société, pouvant inclure des pertes en vies humaines, des dommages sociétaux systémiques ou des menaces existentielles	Problèmes de fiabilité	Risques biologiques et chimiques
		Cyberattaques	Perte de contrôle

³ Cette classification est indicative plutôt qu'exhaustive et fournit une caractérisation générale des risques de préjudice. Le préjudice associé à un risque donné dépend des circonstances dans lesquelles il se concrétise. Par exemple, les dommages liés à une cyberattaque utilisant l'IA sont généralement réversibles. Cependant, elle pourrait également avoir des conséquences irréversibles, telles que des pertes humaines si des hôpitaux sont pris pour cible. Chaque risque de préjudice doit être évalué au cas par cas. Les modèles à poids ouverts n'ont pas été inclus dans le tableau, car il s'agit d'un risque transversal pouvant entraîner tout type de préjudice lié aux autres risques de préjudice liés à l'IA.

Incertitude quant à l'ampleur des préjudices : le cas des contenus pédopornographiques générés par l'IA et des images intimes non consenties



L'IA a rendu plus facile, moins coûteuse et plus rapide la création de contenus pédopornographiques (CSAM) et d'images intimes non consenties (NCII), notamment de la pornographie « deepfake » réaliste et d'images sexualisées d'enfants générées par l'IA, ciblant principalement les femmes et les filles. Ces technologies sont de plus en plus utilisées pour faciliter le harcèlement sexuel, le grooming, la sextorsion et les abus en ligne, en permettant aux délinquants de générer de grands volumes de contenus exploitants avec un minimum de compétences techniques.

L'ampleur du problème augmente rapidement : le Centre national américain pour les enfants disparus et exploités (NCMEC) a reçu environ 4 700 signalements impliquant l'IA générative en 2023, un chiffre qui est passé à 67 000 en 2024, soit une augmentation de 1 325 %. En 2024-2025, le NCMEC a identifié plus de 275 victimes directes de matériel pédopornographique généré par l'IA.

Les préjudices sont graves, que l'image représente un abus réel ou qu'elle soit générée par l'IA. Pour les enfants, le matériel pédopornographique généré par l'IA peut faciliter le « grooming », la coercition, le chantage sexuel à des fins financières et le harcèlement, tout en causant des dommages psychologiques durables, de la honte, de la peur et une perte de contrôle sur son identité pouvant conduire au suicide. De même, les images intimes non consenties (NCII) générées par l'IA infligent des préjudices psychologiques, sociaux et économiques considérables aux adultes et aux adolescents, notamment

l'humiliation, l'atteinte à la réputation, le harcèlement, la perte d'opportunités éducatives ou professionnelles et la crainte pour leur sécurité personnelle. Une enquête menée en 2024 dans dix pays a révélé que 2,2 % des personnes interrogées avaient été confrontées à des images intimes synthétiques non consenties.



Comme les images à caractère sexuel générées par l'IA peuvent être rapidement diffusées, copiées et rester en ligne indéfiniment, elles causent des préjudices durables dont les victimes ont du mal à se libérer.

Bien qu'il existe des statistiques sur l'augmentation du nombre de signalements de CSAM et de NCII générés par l'IA, on manque encore de données fiables sur leur prévalence, leur fréquence et les préjudices qui y sont associés. Les données disponibles reposent encore largement sur les incidents signalés, les détections par les plateformes et les bases de données d'incidents, plutôt que sur des mesures systématiques à l'échelle de la population. De plus, les victimes peuvent être réticentes ou incapables de signaler les incidents en raison de la stigmatisation, de la crainte d'une nouvelle victimisation ou de préoccupations liées à la vie privée et à l'atteinte à la réputation. En conséquence, une grande incertitude subsiste quant à l'ampleur du contenu pédopornographique et des images non consenties générés par l'IA, ainsi qu'à la mesure dans laquelle les mécanismes existants de détection et de signalement permettent de cerner le problème.

La mise en place de voies multiples et complémentaires pour détecter, signaler et mesurer ces incidents pourrait contribuer à renforcer la base de données factuelles et à réduire l'incertitude au fil du temps. Les témoignages des victimes et les échanges avec des associations non gouvernementales pourraient également aider les analystes de politiques à comprendre l'ampleur de l'impact psychologique sur les victimes.

Hiérarchisation des risques

Les risques liés à l'IA dépendent fortement du contexte et varient selon les cas d'utilisation. De nombreux secteurs, tels que la santé, l'industrie automobile et les services financiers, disposent déjà de cadres réglementaires et de gestion des risques solides, conçus pour faire face aux préjudices dans leurs domaines respectifs. Les efforts d'élaboration des politiques devraient donc se concentrer sur les domaines où les cadres existants sont insuffisants ou où l'IA introduit de nouveaux risques qui ne sont pas traités de manière adéquate. Cela nécessite de dresser un état des lieux du paysage des politiques existantes afin d'identifier les lacunes, les législations et réglementations à modifier, ainsi que les domaines où de nouvelles interventions publiques sont nécessaires.

En outre, de nombreuses entreprises spécialisées dans l'IA ont mis en place des mesures de gouvernance pour atténuer les risques associés à leurs modèles, telles que des évaluations de sécurité, des exercices de « red teaming » et des restrictions d'utilisation. Les analystes de politiques devraient évaluer si ces initiatives offrent un niveau de protection adéquat ou si une intervention gouvernementale est nécessaire pour établir des critères de référence cohérents, volontaires ou obligatoires, à l'échelle du secteur.

Exemple de hiérarchisation des risques

Certains cadres de gouvernance existants, tels que les réglementations relatives aux infrastructures critiques, les exigences de « sécurité dès la conception » et de « sécurité par défaut », les tests d'intrusion obligatoires, les exercices de simulation d'attaques (« red teaming »), les audits de sécurité indépendants et les exigences en matière de cybersécurité de la chaîne d'approvisionnement, peuvent déjà s'avérer suffisants pour atténuer le risque de cyberattaques basées sur l'IA contre les infrastructures critiques. Si ce n'est pas le cas, des modifications ciblées ou de nouvelles interventions publiques pourraient s'avérer nécessaires.

Questions guides pour évaluer la nécessité d'une intervention publique

- Quel est le préjudice potentiel, quelle pourrait être la gravité de ses conséquences, et celles-ci seraient-elles réversibles?
- Qui pourrait être touché et quel est son niveau de vulnérabilité?
- Que savons-nous et quelles sont les lacunes en matière de données concernant le risque, ses causes potentielles et ses conséquences? En quoi cela influe-t-il sur la caractérisation du risque?
- Quelles données supplémentaires permettraient d'améliorer la prise de décision et comment les analystes de politiques pourraient-ils les obtenir?
- Quelles hypothèses sous-tendent notre compréhension du risque et dans quelle mesure avons-nous confiance en ces hypothèses?
- À quel niveau ce risque apparaît-il (technique, individuel, sociétal, international) et pourrait-il se propager d'un niveau à l'autre?
- Dans quel délai ce risque se concrétiserait-il et de combien de temps disposons-nous pour intervenir avant que les conséquences ne deviennent difficiles à prévenir ou à atténuer?
- Quelles seraient les conséquences de l'inaction, y compris si la compréhension actuelle du risque s'avérerait correcte?

Options de politiques : définir l'objectif, sélectionner et comparer les mesures

Une fois les risques de préjudice identifiés, analysés et classés par ordre de priorité, les analystes de politiques doivent déterminer leurs objectifs avant de sélectionner les instruments pertinents pour réduire soit la probabilité de concrétisation d'un risque, soit la gravité de ses conséquences, soit la propagation des préjudices, soit la vulnérabilité de la société face à ces préjudices. À ce stade, aucune décision quant à l'opportunité d'intervenir n'a encore été prise. Les analystes de politiques étudient les options permettant de traiter le problème identifié, y compris l'option du statu quo qui sert de référence pour toutes les autres options.

Définir l'objectif des politiques

En situation d'incertitude, les analystes de politiques doivent trouver un équilibre entre la recherche du bien-être social attendu et le renforcement de la résilience sociale. Plutôt que d'évaluer les interventions publiques uniquement en fonction de leurs bénéfices nets attendus, les analystes de politiques devraient également évaluer dans quelle mesure une intervention publique renforce la capacité de la société à se préparer, à résister, à réagir et à s'adapter. À l'instar de la diversification d'un portefeuille financier, la combinaison de plusieurs instruments de politiques réduit la dépendance vis-à-vis d'une mesure unique, améliore la résilience si les hypothèses s'avèrent erronées et crée de multiples lignes de défense contre des risques en constante évolution.

Exemple d'objectif politique

Réduire de 50 % le risque de cyberattaques basées sur l'IA contre les infrastructures critiques au cours des trois prochaines années.

Sélection de mesures de politiques

Les interventions publiques doivent donc être envisagées à travers les différents niveaux décrits dans la section précédente et adaptées au contexte juridique, économique, politique et culturel du pays. Différents instruments remplissent différentes fonctions de gouvernance : certains génèrent des données factuelles et réduisent l'incertitude, d'autres améliorent la sécurité et la fiabilité des systèmes d'IA, renforcent la résilience sociétale ou limitent l'exposition à des préjudices potentiellement irréversibles. Le choix des instruments de politiques doit être guidé par les objectifs de politiques et leur contribution attendue à l'approche de politique globale. Les analystes de politiques peuvent concevoir des mesures à l'aide d'un éventail d'instruments :

→ **Instruments juridiques contraignants** : les instruments juridiques contraignants établissent des exigences obligatoires qui régissent le développement, le déploiement et l'utilisation des systèmes d'IA. Ils comprennent la législation, la réglementation, les exigences en matière d'autorisation et d'autres obligations juridiques exécutoires qui définissent les droits, les responsabilités et les sanctions en cas de non-respect. Ces instruments offrent une sécurité juridique aux entreprises et prévoient la mise en place d'une autorité chargée de veiller au respect des obligations prescrites.

Exemple : la [loi européenne sur l'IA](#) établit un cadre juridiquement contraignant, fondé sur les risques, pour les systèmes d'IA. Elle interdit certaines pratiques inacceptables en matière d'IA, impose des exigences obligatoires aux systèmes d'IA à haut risque et introduit des obligations relatives à la transparence, à l'évaluation de la conformité, à l'étiquetage de certains contenus générés par l'IA et à la surveillance post-commercialisation. Plutôt que de créer des obligations verticales par secteur, elle fournit un cadre global qui s'applique de manière

horizontale à tous les secteurs.

Les « [Mesures provisoires relatives à la gestion des services d'intelligence artificielle générative](#) » de la Chine (2023) exigent le dépôt d'un algorithme et une évaluation de sécurité couvrant la sécurité de l'algorithme, des données et du contenu avant le déploiement de certains services d'IA générative destinés au grand public, faisant ainsi de l'examen réglementaire et de l'enregistrement une condition d'accès au marché.

→ **Instruments de droit souple** : les instruments de droit souple influencent les comportements sans imposer d'obligations juridiquement contraignantes. Ils comprennent des normes volontaires, des spécifications techniques, des codes de conduite, des principes éthiques et des orientations réglementaires qui favorisent les bonnes pratiques, facilitent la conformité réglementaire et soutiennent des approches cohérentes entre les différentes juridictions. Contrairement aux instruments juridiques contraignants, leur flexibilité leur permet généralement d'évoluer au rythme des technologies en mutation rapide, ce qui explique pourquoi ils sont souvent combinés.

Exemple : élaboré par l'Institut national américain des normes et des technologies (NIST), le cadre de gestion des risques liés à l'IA (AI Risk Management Framework) est un cadre volontaire qui aide les organisations à identifier, évaluer et gérer les risques liés à l'IA. Plutôt que de créer des obligations juridiques, il fournit une terminologie commune, des pratiques de gouvernance et des processus de gestion des risques que les organisations peuvent adapter à leur contexte.

→ **Instruments expérimentaux :** les instruments expérimentaux permettent aux régulateurs et aux innovateurs de tester de nouvelles technologies et approches réglementaires dans des conditions contrôlées. Cela inclut les bacs à sable réglementaires, les bancs d'essai, les programmes pilotes, la surveillance, les systèmes de signalement des incidents, les évaluations indépendantes volontaires et les programmes d'aide aux victimes. Ces mécanismes génèrent des données factuelles, réduisent l'incertitude réglementaire et permettent aux cadres de gouvernance d'évoluer en réponse aux changements technologiques.

Exemple : l'Autorité de développement des médias et des technologies de l'information (IMDA) de Singapour et la fondation AI Verify ont mis en place le « Generative AI Evaluation Sandbox » afin de permettre aux développeurs, aux déployeurs et aux testeurs indépendants d'évaluer les applications d'IA générative dans un environnement réel contrôlé. Ce bac à sable favorise l'apprentissage réglementaire, élabore des méthodologies d'évaluation communes et alimente l'élaboration de nouvelles normes d'assurance en matière d'IA.

→ **Instruments programmatiques :** les instruments programmatiques s'appuient sur des programmes et des ressources gouvernementaux pour faire face aux risques liés à l'IA. Ils comprennent notamment le financement public de la recherche et de la sécurité en matière d'IA, les investissements dans les infrastructures de test et d'évaluation, ainsi que les initiatives en faveur de la littératie, de la mise à niveau des compétences et de la reconversion professionnelle. Plutôt que d'imposer des obligations, ces instruments orientent l'innovation en créant des incitations, en renforçant les capacités et en orientant les ressources publiques vers les objectifs de politiques.

Exemple : le programme finlandais « Elements of AI » est un cours en ligne gratuit d'initiation à l'IA développé par l'université d'Helsinki et Reaktor, dont l'expansion européenne est financée par le gouvernement finlandais. Ce programme renforce les capacités du public en investissant dans l'éducation à l'IA et les compétences de la main-d'œuvre, démontrant ainsi comment les gouvernements peuvent promouvoir une IA digne de confiance grâce au renforcement des capacités et à l'investissement public. De même, le Canada a annoncé une initiative nationale de littératie de l'IA dans le cadre de sa stratégie nationale sur l'IA dévoilée en juin 2026. Le Brésil soutient également un programme de vérification des faits qui permet à des tiers de vérifier de manière indépendante les informations et de contribuer à lutter contre la désinformation générée par l'IA.

→ Instruments d'incitation économique :

Les instruments d'incitation économique visent à influencer le développement et le déploiement de l'IA en modifiant les incitations financières auxquelles sont confrontés les développeurs et les déployeurs, plutôt qu'en imposant des obligations légales directes. Ils comprennent les marchés publics, les subventions et aides financières, les prix, les incitations fiscales, les mécanismes d'assurance et d'autres outils financiers qui encouragent l'investissement dans une IA fiable et sûre. En récompensant les pratiques souhaitables plutôt qu'en les imposant, ces instruments peuvent compléter les approches réglementaires tout en soutenant l'innovation.

Exemple : le mémorandum du Bureau de la gestion et du budget des États-Unis intitulé « Driving Efficient Acquisition of Artificial Intelligence in Government » (2025) établit des exigences à l'échelle de l'administration fédérale pour les marchés publics liés à l'IA passés par les agences fédérales. Il enjoint aux agences de tester, dans la mesure du possible, les solutions d'IA proposées afin d'évaluer leurs capacités et leurs limites, y compris sur les systèmes appartenant à l'État. Il exige également que les agences conservent la capacité de surveiller et d'évaluer les performances et les risques des systèmes d'IA déployés. En faisant de ces pratiques des conditions préalables aux marchés publics fédéraux, cette politique publique utilise le pouvoir d'achat de l'État pour encourager le développement de systèmes d'IA fiables, sécurisés et dignes de confiance.

→ Instruments institutionnels :

Les instruments institutionnels renforcent l'écosystème de gouvernance en mettant en place les organisations, les capacités techniques, les processus et les accords de collaboration nécessaires pour comprendre et gérer efficacement les risques liés à l'IA. Ils comprennent des autorités spécialisées en matière de réglementation et d'évaluation, des mécanismes de coopération nationaux et internationaux, des partenariats public-privé et d'autres dispositifs institutionnels qui contribuent à renforcer les connaissances et les compétences institutionnelles.

Exemple : L'Institut britannique pour la sécurité de l'IA (UK AI Security Institute) est une institution publique spécialisée chargée d'évaluer les modèles d'IA de pointe, d'élaborer des méthodologies d'évaluation de la sécurité et de soutenir une gouvernance de l'IA fondée sur des données probantes. Il collabore également à l'échelle internationale avec des institutions homologues aux États-Unis, au Canada, en France, à Singapour, en Corée du Sud et au Japon par le biais du Réseau international des instituts pour la sécurité de l'IA, illustrant ainsi comment des organismes de gouvernance dédiés peuvent renforcer les capacités réglementaires et la coordination.

Exemple d'options de politiques

1. **Instrument institutionnel** : créer un mécanisme public-privé de partage de l'information et des vulnérabilités afin de partager les failles de sécurité et les renseignements sur les menaces parmi les acteurs du secteur de l'IA.
2. **Instrument de droit souple** : des lignes directrices recommandant des mesures minimales de cybersécurité et des tests de type « red teaming » pour les modèles d'IA afin de résister au « jailbreaking ».
3. **Instrument expérimental** : mener une expérience contrôlée de cyberattaque assistée par l'IA contre des infrastructures critiques afin de renforcer les cyberdéfenses.
4. **Instrument juridique contraignant** : contrôles obligatoires du déploiement visant à restreindre l'accès aux modèles avancés aux clients ayant fait l'objet d'une vérification préalable

Comparaison des options

L'analyse coûts-bénéfices reste un outil important pour comparer les différentes options de politiques, mais comme les estimations quantitatives des coûts et des bénéfices liés aux risques de l'IA restent très incertaines, voire indisponibles, elle doit être adaptée à la réalité spécifique de la gestion des risques en situation d'incertitude. Les analystes de politiques devraient compléter leur analyse coûts-bénéfices par des évaluations qualitatives qui prennent en compte l'efficacité attendue d'une intervention, sa robustesse face à divers scénarios plausibles, sa réversibilité si les hypothèses s'avéraient erronées, sa flexibilité et sa capacité à s'adapter à mesure que les données évoluent, son potentiel à générer de nouvelles connaissances, ses impacts sur l'innovation, sa faisabilité administrative et sa compatibilité avec les approches de gouvernance internationale.

Le statu quo, c'est-à-dire la décision des gouvernements de ne pas intervenir, constitue également une option. Bien que certains milieux de politiques craignent que l'intervention gouvernementale, en particulier la réglementation, puisse entraver l'innovation en matière d'IA, la relation entre réglementation et innovation est plus nuancée. Les « politiques en matière d'IA » englobe un large éventail d'instruments et de types d'intervention dont chacun peut avoir des effets différents sur l'innovation, tant positifs que négatifs. Tout comme pour l'évaluation des préjudices potentiels associés aux risques liés à l'IA, déterminer les effets d'une intervention publique sur l'innovation exige des analystes de politiques qu'ils examinent les données disponibles au cas par cas.

Les analystes de politiques doivent mettre en balance l'intérêt de retarder l'action pour recueillir des données supplémentaires et la gravité potentielle

ainsi que la réversibilité du préjudice s'il venait à se concrétiser. Face à des risques modérés et réversibles, les analystes de politiques peuvent disposer d'un délai plus long pour collecter des données, suivre l'évolution de la situation et adapter de manière itérative les programmes existants ou concevoir de nouvelles interventions publiques. En revanche, des risques critiques et irréversibles peuvent nécessiter une action préventive malgré des données limitées, les politiques étant ensuite ajustées à mesure que les connaissances s'améliorent.

EXEMPLE DE COMPARAISON DES OPTIONS						
Options	Résultat attendu	Coût	Échéancier avant l'impact	Incidence sur le marché	Flexibilité	Efficacité attendue
Statu quo	Poursuite des activités habituelles	Potentielle-ment élevés Dommages liés aux cyberattaques en cas d'incident	Sans objet	Aucun	Élevé Préserve la flexibilité pour de futures interventions	Aucune Aucune réduction de la probabilité ou de l'impact des cyberattaques basées sur l'IA
Mécanisme de partage d'informations sur les vulnérabilités et les menaces	Permettre aux pouvoirs publics et aux entreprises d'identifier, de partager et de réagir plus rapidement	Moyen Nécessite la mise en place d'un mécanisme	Moyen Les avantages augmentent à mesure que la participation et la confiance se renforcent	Modéré	Élevé Le mécanisme peut évoluer	Modérée L'efficacité dépend de la participation du secteur et de la qualité de l'information
Mesures minimales de cybersécurité et lignes directrices pour les exercices de simulation d'attaques	Inciter les entreprises spécialisées dans l'IA à réduire leurs vulnérabilités informatiques	Faible Élaborer et mettre à jour régulièrement des lignes directrices	Court Les entreprises peuvent adopter ces lignes directrices rapidement	Faible Mise en œuvre volontaire	Élevé Les lignes directrices peuvent être mises à jour	Faible L'adoption peut être inégale, ce qui limite la réduction globale des risques
Expérience contrôlée de cyberattaque utilisant l'IA	Générer des données probantes pour améliorer la résilience des infrastructures critiques	Moyen Nécessite une expertise spécialisée et un environnement de test sécurisé	Moyen La planification, la réalisation et l'évaluation de l'exercice prennent du temps	Faible Participation volontaire	Élevée Les scénarios expérimentaux peuvent être adaptés	Modéré Les résultats ne sont pas nécessairement transposables à des attaques réelles
Contrôles de déploiement	Réduire l'accès des acteurs malveillants aux capacités avancées de l'IA	Élevé Nécessite une supervision réglementaire, un suivi et des mesures coercitives	Long Nécessite des autorités judiciaires et un mécanisme d'application de la loi	Élevé Peut avoir une incidence sur l'innovation et la concurrence sur le marché	Modéré Les modifications réglementaires visant à ajuster les contrôles d'accès peuvent être lentes	Élevé Dépend de la solidité des contrôles d'accès et de l'efficacité de leur application

Questions clés pour évaluer les options politiques

- Quel est l'objectif de la politique?
- Quelles sont les options politiques disponibles, y compris le statu quo?
- Quel est le résultat attendu? De quelles données avons-nous besoin pour évaluer s'il a été atteint?
- Quels sont les coûts et les bénéfices de chaque option? Dans quelle mesure sont-ils sensibles à l'incertitude?
- Existe-t-il des risques ou des effets indésirables potentiels?
- Quelle est l'efficacité attendue de l'intervention et quelles sont les hypothèses sous-jacentes?
- Dans quelle mesure cette option est-elle solide face à divers scénarios plausibles?
- La mesure est-elle réversible si les données et les conditions évoluent?
- Dans quels délais la mesure peut-elle être adaptée pour tenir compte de nouvelles données?
- Quels indicateurs ou seuils pourraient signaler la nécessité d'ajuster la politique?
- Quel est le mécanisme de révision de la mesure si elle venait à être adoptée?
- Cette option préserve-t-elle la flexibilité politique future, ou pourrait-elle créer des dépendances vis-à-vis d'une trajectoire donnée?
- La mesure est-elle réalisable et applicable?
- Cette option permet-elle de générer les données nécessaires pour mieux comprendre le risque en question?
- Dans quelle mesure cette mesure interfère-t-elle avec le marché et quel serait son impact sur l'innovation?
- Dans quelle mesure est-elle compatible avec les approches de gouvernance internationales?
- Dans quelle mesure cette option complète-t-elle d'autres mesures déjà en place ou à adopter? Les interactions avec d'autres mesures pourraient-elles amplifier ou atténuer le risque ciblé?
- Quelle est la tolérance au risque de la société⁴ concernant le risque ciblé?

⁴ Selon le cadre de gestion des risques liés à l'IA du NIST, la tolérance au risque est «le degré d'incertitude qu'une société est prête à accepter et à supporter. Le niveau de risque acceptable pour une société dépend fortement du contexte et est susceptible d'évoluer au fil du temps à mesure que de nouvelles données apparaissent».

Exemples d'approches de politiques

En fonction du moment, de la gravité et de la réversibilité du préjudice, de la tolérance au risque d'une société, de sa maturité institutionnelle et de ses contextes juridiques, économiques, politiques et culturels, les analystes de politiques peuvent adopter diverses interventions, débouchant sur des approches de politiques uniques. Il convient de noter que le degré d'incertitude ne détermine pas l'approche de politique. Ce que l'on sait des caractéristiques et des conséquences potentielles du préjudice détermine la position politique initiale; l'incertitude détermine quant à elle la manière dont les interventions doivent être conçues et gérées au fil du temps. La section suivante présente quatre approches basées sur la matrice de classification des risques liés à l'IA. Ces approches ne s'excluent en aucun cas mutuellement et peuvent être combinées à différents niveaux d'intervention publique afin de créer un portefeuille de politiques diversifié qui reste solide face à l'incertitude.

Axée sur la résilience

Type de préjudice	Horizon temporel	Niveau d'intervention	Type d'instruments politiques	Gestion de l'incertitude	Degré d'ingérence sur le marché	Flexibilité	Coopération internationale
Conséquences modérées et réversibles	À plus long terme	Sociétal	Program-matique Institutionnel	Suivre et adopter des mesures	Faible	Élevé	Échange de connaissances et de bonnes pratiques

Une approche axée sur la résilience est la plus appropriée pour les risques modérés et réversibles, car ceux-ci apparaissent généralement de manière progressive et se développent sur le long terme, comme la désinformation facilitée par l'IA, l'érosion de l'esprit critique, les perturbations du marché du travail et les violations des droits fondamentaux. L'objectif est de profiter de cette opportunité pour collecter des données afin de mieux comprendre le risque et de renforcer la capacité des individus, des organisations et des institutions publiques à s'adapter aux perturbations. Les mesures pertinentes relèvent principalement des instruments programmatiques et du renforcement des capacités institutionnelles; elles comprennent notamment des initiatives publiques en matière d'IA et de littératie numérique, des programmes de perfectionnement et de reconversion professionnelle des travailleurs, un accompagnement à la transition de la main-d'œuvre, des observatoires de recherche, ainsi que l'adaptation des cadres réglementaires relatifs aux élections, à la protection des consommateurs et à la protection des données. Cette approche est très flexible et préserve les options futures, car les mesures peuvent être affinées en fonction des nouvelles données disponibles. Bien qu'elle soit principalement mise en œuvre au niveau national, la coopération internationale facilite l'échange de connaissances et de bonnes pratiques.

Expérimental

Type de préjudice	Horizon temporel	Niveau d'intervention	Type d'instruments politiques	Gestion de l'incertitude	Degré d'ingérence sur le marché	Flexibilité	Coopération internationale
Significative mais réversible et modérée, irréversible	À court terme	Modèle et individu	Expérimental Program-matique	Produire des données probantes et ajuster les mesures	Faible	Élevé	Partage d'informations et méthodologies d'évaluation communes

Une approche expérimentale est particulièrement adaptée aux risques modérés, irréversibles et aux préjudices importants mais réversibles, et lorsque les données factuelles relatives à un risque ne peuvent être générées qu'après un déploiement en conditions réelles. Citons par exemple la fraude assistée par l'IA, la production et la diffusion de contenus pédopornographiques, ainsi que les images intimes non consenties. Dans de tels cas, les analystes de politiques devraient combiner des mesures visant à atténuer les préjudices existants avec des mécanismes permettant de générer des données factuelles sur les schémas d'utilisation abusive et l'ampleur des préjudices. Au-delà de l'adaptation et de l'application des lois existantes, des instruments expérimentaux tels que les « bacs à sable » réglementaires, la surveillance et les systèmes de signalement d'incidents peuvent être associés à des mesures programmatiques, comme les programmes d'aide aux victimes. Cette approche permet aux analystes de politiques de tirer des enseignements de l'expérience concrète sans imposer prématurément des exigences rigides lorsque la base de données probantes reste incomplète. Bien que la mise en œuvre relève principalement de la compétence nationale, la coopération internationale peut renforcer l'apprentissage collectif grâce au partage d'informations, à des cadres de signalement communs et à des méthodologies d'évaluation partagées.

Basée sur l'assurance

Type de préjudice	Horizon temporel	Niveau d'intervention	Type d'instruments politiques	Gestion de l'incertitude	Degré d'ingérence sur le marché	Flexibilité	Coopération internationale
Critique mais réversible et significatif, irréversible	À court terme	Modèle	Juridique contraignant Droit non contraignant	Demander des preuves et adapter les mesures	Moyen	Moyen	Critères d'évaluation communs, interopérables ou mutuellement reconnus

L'approche fondée sur l'assurance est la plus appropriée pour les préjudices critiques mais réversibles et les préjudices irréversibles significatifs, lorsque des preuves peuvent être obtenues par des tests et des évaluations préalables au développement, tels que des exercices de « jailbreaking » ou des tests d'alignement. L'objectif est de démontrer que les systèmes d'IA répondent à un ensemble d'exigences de sécurité avant et pendant leur déploiement, en exigeant des entreprises qu'elles fournissent des preuves à l'appui de leurs allégations en matière de sécurité. Les instruments de politiques peuvent inclure des évaluations de risques obligatoires, des évaluations de sécurité, des audits indépendants, des normes techniques, la certification, des exigences en matière de documentation, la déclaration des incidents, l'octroi de licences et la surveillance réglementaire. Les instruments juridiques contraignants peuvent établir des exigences de sécurité de haut niveau et des obligations de responsabilité, tandis que les instruments de droit souple fournissent des normes, des référentiels d'évaluation, des orientations et des méthodologies susceptibles d'évoluer à mesure que de nouvelles preuves apparaissent. Cette combinaison permet aux analystes de politiques d'établir une base réglementaire stable tout en conservant une flexibilité suffisante pour adapter la mise en œuvre à mesure que la compréhension des risques et de la technologie s'améliore. Bien que plus contraignante que l'approche expérimentale, elle définit un cadre de responsabilité pour les entreprises du secteur de l'IA et renforce la confiance des analystes de politiques quant à la mise en place de garanties appropriées. La coordination internationale peut réduire les doublons en favorisant des normes communes, la reconnaissance mutuelle des évaluations de conformité et des cadres d'évaluation interopérables.

Préventif

Type de préjudice	Horizon temporel	Niveau d'intervention	Type d'instruments politiques	Gestion de l'incertitude	Degré d'ingérence sur le marché	Flexibilité	Coopération internationale
Critique et irréversible	À long terme	Modèle	Juridique contraignant	Mettre en place des mesures et assurer un suivi	Élevé	Faible	Obligations de sécurité de base communes

Lorsque les risques potentiels sont critiques et irréversibles, les analystes de politiques peuvent être amenés à agir malgré une incertitude importante, plutôt que d'attendre de disposer de preuves suffisantes pour établir avec certitude leur probabilité ou leur ampleur. On peut citer, à titre d'exemple, les dommages irréversibles pour la santé publique résultant d'une attaque biologique ou chimique orchestrée par l'IA, ou encore les pertes humaines massives causées par une cyberattaque menée par une IA agentique contre des infrastructures critiques. Dans de telles circonstances, la prévention de dommages catastrophiques devrait primer sur la préservation des avantages potentiels liés à l'innovation. Les mesures peuvent inclure des seuils de capacité, des restrictions de déploiement, des exigences de sécurité renforcées, des interdictions de réplique autonome ou de poursuite d'objectifs, ainsi que des restrictions sur les applications particulièrement dangereuses jusqu'à ce que des preuves suffisantes de leur sécurité aient été apportées. Bien que les mesures de précaution puissent freiner l'innovation à court terme pour certaines capacités ciblées de l'IA, elles réduisent l'exposition à des dommages irréversibles tout en laissant le temps à la compréhension scientifique, aux capacités de gouvernance et aux techniques de sécurité de l'IA de mûrir. Cette approche nécessite un haut degré de coordination internationale pour respecter des obligations de sécurité de base communes; si les pays agissaient dans leur propre intérêt national face à un risque critique et irréversible, cela pourrait conduire à une issue catastrophique pour tous.

Arbre de décision pour la gestion des risques liés à l'IA en situation d'incertitude

1. **Classer le préjudice potentiel** (c'est-à-dire gravité, réversibilité, délai)
2. **Déterminer l'approche politique appropriée** (par exemple, axée sur la résilience, expérimentale, fondée sur l'assurance, préventive ou toute autre)
3. **Identifier les incertitudes qui ont une incidence sur la décision** (par exemple : ce que nous ignorons, l'importance de ces inconnues et les hypothèses susceptibles de modifier la décision)
4. **Concevoir l'intervention pour gérer cette incertitude** (par exemple : surveiller, générer des données probantes, demander des garanties, identifier les déclencheurs d'ajustement, prévoir une marge de manœuvre)
5. **Réévaluer à mesure que les données probantes et les circonstances évoluent** (c'est-à-dire prévoir un mécanisme de révision)

Conclusion

Face à l'incertitude entourant les risques liés à l'IA, les analystes de politiques devraient viser à renforcer la résilience sociétale en réduisant les vulnérabilités, en limitant la propagation des préjudices et en préservant la capacité d'adaptation à mesure que les connaissances et les conditions évoluent. Cette boîte à outils s'appuie sur la liste non exhaustive des risques émergents liés à l'IA et des lacunes en matière de données associées identifiées dans le [Rapport international sur la sécurité de l'IA](#). Elle fournit un cadre permettant d'adapter le processus d'élaboration des politiques à l'incertitude en s'appuyant sur des approches établies de gestion des risques et de prise de décision ainsi que sur les enseignements tirés d'autres domaines critiques en matière de sécurité. Bon nombre de ces enseignements sont repris dans le [Rapport international sur la sécurité de l'IA](#) et dans le [Cadre de gestion des risques liés à l'IA du NIST](#). Ils peuvent être résumés en cinq points :

1. **Premièrement, les évaluations de risques doivent être ré-examinées régulièrement.** Les analystes de politiques doivent actualiser leur compréhension de la manière dont les risques pourraient se concrétiser, du moment, de la gravité et de la réversibilité des préjudices, ainsi que des hypothèses qui sous-tendent leurs évaluations.
2. **Deuxièmement, l'attention des analystes de politiques doit se concentrer sur les lacunes que les cadres existants ne parviennent pas à combler de manière adéquate.** Cela inclut les domaines où la réglementation existante et la gouvernance volontaire du secteur sont insuffisantes, ainsi que les risques inédits introduits ou amplifiés par l'IA.
3. **Troisièmement, les objectifs de politiques doivent donner la priorité au renforcement de la résilience sociétale à travers quatre niveaux d'intervention :** le modèle, l'individu, la société et le niveau international. Les analystes de politiques doivent choisir des interventions qui renforcent la capacité à se préparer, à résister, à réagir et à s'adapter aux préjudices liés à l'IA, tout en adoptant des mesures visant à combler le déficit de données à chaque niveau d'intervention en fonction de la nature de l'incertitude à laquelle ils sont confrontés.
4. **Quatrièmement, l'analyse coûts-bénéfices doit être complétée par d'autres considérations tenant compte de l'incertitude.** Celles-ci comprennent l'efficacité attendue, la robustesse face à des scénarios d'avenir plausibles, la réversibilité, l'adaptabilité, la capacité à générer des connaissances, les effets sur l'innovation, la faisabilité administrative et la compatibilité avec les approches internationales.

5. **Enfin, l'inaction est en soi un choix de politique.** Les analystes de politiques doivent mettre en balance l'intérêt de retarder une intervention afin de recueillir des données supplémentaires et les conséquences potentielles de laisser le risque se concrétiser, en particulier lorsque les préjudices pourraient être graves ou irréversibles.

En résumé, cette boîte à outils offre aux analystes de politiques un nouveau cadre de réflexion pour aborder la gestion des risques liés à l'IA en situation d'incertitude. Elle jette les bases d'une approche plus systématique de la comparaison des interventions politiques en complétant l'analyse coûts-bénéfices par des critères qualitatifs tels que la robustesse, la réversibilité, l'adaptabilité, la production de données et la faisabilité. Ce cadre pourrait être mis en œuvre grâce à la création d'un répertoire évolutif qui filtrerait les mesures politiques en fonction de leurs objectifs, de leur niveau d'intervention, de l'instrument politique utilisé, du délai prévu pour l'impact, du degré d'ingérence sur le marché, de la flexibilité et d'autres considérations pertinentes. Continuellement mis à jour avec les nouvelles données et les pratiques de gouvernance émergentes, une telle ressource pratique pourrait aider les analystes de politiques du monde entier à explorer différentes options politiques avant de mener leur propre analyse politique détaillée.

Car, en fin de compte, l'incertitude ne peut être éliminée avant que les décisions politiques ne soient prises. Mais les politiques peuvent être conçues pour fonctionner efficacement dans un contexte d'incertitude.

Références

1. Acton, M. (2026). « Un agent d'OpenAI a piraté le site web d'un service de santé australien », Financial Times, 23 septembre 2026, <https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals>
2. Agence américaine des produits alimentaires et médicamenteux (2022). Processus de développement et d'autorisation | Médicaments, 8 août 2022, <https://www.fda.gov/drugs/development-approval-process-drugs>
3. Agence internationale de l'énergie atomique (1992). Évaluation probabiliste de la sûreté, série INSAG, 12 juillet 1992, <http://www.iaea.org/publications/3789/probabilistic-safety-assessment>
4. Anthropic (2026). Enquête sur trois incidents réels dans le cadre de nos évaluations de cybersécurité, 30 juillet 2026, <https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals>
5. Bengio, Y. et al. (2026). Rapport international sur la sécurité de l'IA, 3 février 2026, <https://internationalaisafetyreport.org/>
6. Ben-Haim, Y. (2019). « Info-Gap Decision Theory (IG) » dans Marchau, V. A. W. J. et al. La prise de décision en situation de profonde incertitude : de la théorie à la pratique, Springer International Publishing, pp. 93-115, https://doi.org/10.1007/978-3-030-05252-2_5
7. Bureau des Nations Unies pour la réduction des risques de catastrophe (2021), Modélisation déterministe et probabiliste des risques, 9 juin 2021, <https://www.preventionweb.net/understanding-disaster-risk/key-concepts/deterministic-probabilistic-risk>
8. Bureau du commissaire à l'information (2024). Annexe 5 : Dossiers de justification fondés sur des arguments, 19 novembre 2024, <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/annexe-5-argument-based-assurance-cases/>
9. Chui M. et al. (2023), « Le potentiel économique de l'IA générative », McKinsey & Company, juin 2023, <https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier>
10. De Groot, K., & Thurik, R. (2018). « Disentangling Risk and Uncertainty : When Risk-Taking Measures Are Not About Risk », Frontiers in Psychology, vol. 9, 14 novembre 2018, <https://doi.org/10.3389/fpsyg.2018.02194>
11. Desilva, M. (2024). Qu'est-ce que l'analyse en nœud papillon dans la gestion des risques ?, Riskonnect, 10 janvier 2024, <https://riskonnect.com/reporting-analytics/what-is-bowtie-analysis-in-risk-management/>
12. Fjader, C. (2020). Qu'est-ce que l'« incertitude » et en quoi diffère-t-elle du « risque » ?, Institut de gestion stratégique des risques, 6 novembre 2020, <https://www.theismr.org/what-is-uncertainty-and-how-it-differs-from-risk/>
13. Flinders, C., Barnhart, B., Morrison, E. B., Anderson, P. D. et Landis, W. G. (2025). « Approches probabilistes pour l'évaluation des risques et l'élaboration de critères réglementaires : applications actuelles, lacunes et opportunités », Integrated Environmental Assessment and Management, vol. 21, n° 6, novembre 2025, p. 1281-1292. <https://doi.org/10.1093/inteam/vjaf016>

14. Flynn, S. (2019). Analyse des risques et politiques publiques, EBSCO, 2019, <https://www.ebsco.com/research-starters/business-and-management/risk-analysis-and-public-policy>
15. Fortinet. Qu'est-ce que la défense en profondeur ? Définition et explication, <https://www.fortinet.com/resources/cyberglossary/defense-in-depth>
16. Fortinet. Qu'est-ce que la modélisation des menaces ? Comment fonctionne-t-elle ?, <https://www.fortinet.com/resources/cyberglossary/threat-modeling>
17. Garibian, E. (2025). Qu'est-ce que l'analyse des modes de défaillance et de leurs effets (AMDE) ?, LLumin, 15 janvier 2025, <https://llumin.com/blog/what-is-failure-mode-and-effect-analysis/>
18. Grey, M. et C-R Segerie, Le manuel libre sur la sécurité de l'IA, Centre français pour la sécurité de l'IA, <https://ai-safety-atlas.com/>
19. Groupe scientifique international indépendant des Nations Unies sur l'IA (2026). « Rapport préliminaire », 1er juillet 2026, <https://www.un.org/independent-international-scientific-panel-ai/en/preliminary-report>
20. Guerra, S. et al., « Élaboration d'une stratégie de biosécurité en profondeur pour l'ère de l'IA », Rand, 18 août 2026, https://www.rand.org/pubs/research_reports/RRA4999-1.html
21. Haas, E. J. et Yorio, P. L. (2018). Utilisation de la théorie de la construction du sens pour améliorer la gestion et la communication des risques : que pouvons-nous en tirer ?, Centre de contrôle des maladies, 5 novembre 2018, <https://stacks.cdc.gov/view/cdc/227824>
22. Haasnoot, M., Warren, A. et Kwakkel, J. H. (2019). « Dynamic Adaptive Policy Pathways (DAPP) » dans Marchau, V. A. W. J. et al. La prise de décision en situation d'incertitude profonde : de la théorie à la pratique, Springer International Publishing, pp. 71-92, https://doi.org/10.1007/978-3-030-05252-2_4
23. Hadjisotiriou, S., Marchau, V., Walker, W. et Rikkert, M. O. (2023). « Prise de décision en situation d'incertitude profonde pour la planification des politiques en cas de pandémie », Health Policy, vol. 133, juillet 2023, <https://doi.org/10.1016/j.healthpol.2023.104831>
24. Institut national des normes et des technologies (2023). Cadre de gestion des risques liés à l'intelligence artificielle (AI RMF 1.0), <https://doi.org/10.6028/NIST.AI.100-1>
25. Kasperson, R. E., Webler, T., Ram, B., & Sutton, J. (2022). « Le cadre de l'amplification sociale du risque : nouvelles perspectives ». Risk Analysis, vol. 42, n° 7, 20 juillet 2022, p. 1367-1380. <https://doi.org/10.1111/risa.13926>
26. Kay, J. et M. King (2020). Incertitude radicale : la prise de décision au-delà des chiffres, W.W. Norton & Company, 2020, 527 p.
27. Khakzad, N., Khan, F., Amyotte, P. et Cozzani, V. (2013). « Risk Management of Domino Effects Considering Dynamic Consequence Analysis », Risk Analysis, vol. 34, n° 6, 31 décembre 2013, p. 1128-1138. <https://doi.org/10.1111/risa.12158>
28. Labonte, M. (2023), Introduction aux services financiers : le risque systémique, Service de recherche du Congrès, 1er mai 2023, <https://www.congress.gov/crs-product/IF10700>
29. Lempert, R. J. (2019). « Robust Decision Making (RDM) » dans Marchau, V. A. W. J. et al., Decision Making under Deep Uncertainty : From Theory to Practice, Springer International Publishing, pp. 23-51, https://doi.org/10.1007/978-3-030-05252-2_2
30. Madani, P. (2026). « Comparaison des réglementations en matière d'IA : UE, États-Unis, Royaume-Uni, Chine (2026) », Legalithm, 7 avril 2026, <https://www.legalithm.com/en/blog/>

31. Maison Blanche (2026). Promouvoir l'innovation et la sécurité en matière d'intelligence artificielle avancée, décret présidentiel, 2 juin 2026, <https://www.whitehouse.gov/presidential-actions/2026/06/promoting-advanced-artificial-intelligence-innovation-and-security/>
32. Marchau, V. A. W. J. et al. (2019). La prise de décision en situation d'incertitude profonde : de la théorie à la pratique, Springer International Publishing, <https://doi.org/10.1007/978-3-030-05252-2>
33. Myhr, A. I., & Traavik, T. (2002). « Le principe de précaution : incertitude scientifique et recherches omises dans le contexte de l'utilisation et de la dissémination des OGM », Journal of Agricultural and Environmental Ethics, vol. 15, mars 2002, pp. 73-86, <https://www.salmone.org/wp-content/uploads/2013/09/myhr-2002.pdf>
34. Narayanan, A. et S. Kapoor (2025). Les probabilités de risques existentiels liés à l'IA sont trop peu fiables pour éclairer les politiques, Substack, 15 avril 2025, <https://www.normaltech.ai/p/ai-existential-risk-probabilities>
35. Nations Unies (2026). « Les principaux facteurs de risque de perte de contrôle de l'IA se sont conjugués lors de l'incident de 2026, selon un groupe d'experts scientifiques indépendants de l'ONU », Communiqué de presse, 21 septembre 2026, https://www.un.org/independent-international-scientific-panel-ai/sites/default/files/2026-09/Communiqué%20de%20presse_Note%20thématique_Agents%20d'IA,%20désalignement%20et%20risque%20de%20perte%20du%20contrôle%20humain_Groupe%20scientifique%20sur%20l'IA.pdf
36. Groupe d'experts scientifiques international indépendant des Nations Unies sur l'IA (2026). « Agents d'IA, désalignement et risque de perte de contrôle humain : enseignements tirés de l'incident OpenAI-Hugging Face », Note, version 1, septembre 2026, https://www.un.org/independent-international-scientific-panel-ai/sites/default/files/2026-09/Thematic%20Brief_AI%20Agents%20Misalignment%20and%20the%20Risk%20of%20Losing%20Human%20Control_Evidence%20from%20the%20OpenAI-Hugging%20Face%20Incident_Independent%20International%20Scientific%20Panel%20on%20AI_Advance%20Unedited%20Version%201_21%20sept%202026.pdf
37. OCDE (2021). « Outils pour une IA digne de confiance », Documents de l'OCDE sur l'économie numérique, 28 juin 2021, https://www.oecd.org/en/publications/tools-for-trustworthy-ai_008232ec-en.html
38. OCDE (2024). « Évaluation des risques, des avantages et des impératifs politiques potentiels liés à l'intelligence artificielle », Documents de l'OCDE sur l'intelligence artificielle. <https://doi.org/10.1787/3f4e3dfb-en>
39. OCDE (2025). « Un outil de cartographie des cadres réglementaires numériques, comprenant un projet pilote sur les efforts de réglementation de l'IA », Documents de travail de l'OCDE sur la politique réglementaire, 12 mai 2025, https://www.oecd.org/en/publications/a-map-ping-tool-for-digital-regulatory-frameworks_1cdad902-en.html?utm_source=chatgpt.com
40. OCDE (2026). « Exploration des trajectoires possibles de l'IA jusqu'en 2030 », Document de l'OCDE sur l'intelligence artificielle, <https://doi.org/10.1787/cb41117a-en>
41. OCDE (2026), « L'utilisation de l'IA par les particuliers connaît une forte hausse dans l'ensemble de l'OCDE, tandis que son adoption par les entreprises continue de s'étendre », Communiqué, 28 janvier 2026, <https://www.oecd.org/fr/about/news/announcements/2026/01/>

[ai-use-by-individuals-surges-across-the-oecd-as-adoption-by-firms-continues-to-expand.html](#)

42. OpenAI (2026). OpenAI et Hugging Face unissent leurs forces à la suite d'incidents de sécurité survenus lors de l'évaluation d'un modèle, 21 juillet 2026, <https://openai.com/index/hugging-face-model-evaluation-security-incident/>
43. Organisation de l'aviation civile internationale (2018). « Chapitre 9 : Systèmes de gestion de la sécurité (SMS) » dans le Manuel de gestion de la sécurité (Doc 9859), <https://www.icao.int/safety-management/SMI/SMM/Chapter%209>
44. Patient Safety Network (2024). Analyse des causes profondes, 15 juin 2024, <https://psnet.ahrq.gov/primer/root-cause-analysis>
45. Pimentel, B. (2024). Évaluation des risques : aperçu général, blog juridique de Thomson Reuters, 4 juin 2024, <https://legal.thomsonreuters.com/blog/what-is-a-risk-assessment/>
46. Reuters (2026). « Le modèle d'IA de la start-up chinoise Moonshot s'échappe de son environnement de test, selon des chercheurs », Reuters, 7 août 2026, <https://www.reuters.com/legal/litigation/chinese-startup-moonshots-ai-model-breaks-out-testing-environment-researchers-2026-08-07/>
47. Savin, A. (2026). « Quand le droit se heurte à l'inconnu : risque, incertitude et limites de la réglementation », European Journal of Risk Regulation, juin 2026, <https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/when-the-law-meets-the-unknown-risk-uncertainty-and-the-limits-of-regulation/8B57E1F99088250B250E893A980E4A2E>
48. Simpson, M., A. Ortega et R. Trager (2025). « Initiatives volontaires du secteur privé dans la gouvernance de l'IA de pointe : leçons tirées de l'aviation et de l'énergie nucléaire », 20 janvier 2025, <https://www.oxfordmartin.ox.ac.uk/publications/voluntary-industry-initiatives-in-frontier-ai-governance-lessons-from-aviation-and-nuclear-power>
49. Taleb, N. N. et al. (2014). Le principe de précaution (avec application à la modification génétique des organismes), série de documents de travail de la NYU School of Engineering, 4 septembre 2014, <https://www.fooledbyrandomness.com/pp2.pdf>
50. Valera, F. (2025). Une gestion efficace des risques dans l'industrie automobile, Visure Solutions, 10 avril 2025, <https://visuresolutions.com/automotive/risk-management/>
51. Walker, W. E., Marchau, V. A. W. J. et Kwakkel, J. H. (2019). « Dynamic Adaptive Planning (DAP) » dans Marchau, V. A. W. J. et al., La prise de décision en situation d'incertitude profonde : de la théorie à la pratique (p. 53-69), Springer International Publishing, https://doi.org/10.1007/978-3-030-05252-2_3
52. Wiegmann, D. A. et al. (2022). « Comprendre le « modèle du fromage suisse » et son application à la sécurité des patients », Journal of patient safety, vol. 18, n° 2, mars 2022, p. 119-123. <https://doi.org/10.1097/PTS.0000000000000810>

Annexe I

Fiches de risques

CONTENU GÉNÉRÉ PAR L'IA ET CRIMINALITÉ

Catégorie de risque

Risque lié à une mauvaise utilisation

Risque lié à un dysfonctionnement

Risque systémique

Description

Utilisation de l'IA pour créer du contenu textuel, audio et vidéo réaliste à des fins de fraude, d'escroquerie, de chantage, d'abus, de harcèlement et d'extorsion. Cela inclut les contenus pédopornographiques et les images intimes non consenties qui ciblent de manière disproportionnée les femmes et les filles.

Probabilité

Il existe déjà des cas avérés de criminels utilisant des clones vocaux générés par l'IA et des « deepfakes » pour usurper l'identité de dirigeants ou de membres de la famille afin d'escroquer leurs victimes. De même, des systèmes d'IA ont été utilisés pour générer du matériel pédopornographique et des images intimes non consenties. Comme ces pratiques existent déjà, la probabilité est élevée.

Gravité de l'impact

L'impact de la criminalité facilitée par l'IA se fait principalement sentir au niveau individuel ou au niveau des entreprises, mais s'amplifie à mesure que le nombre de victimes augmente:

- Pertes financières
- Usurpation d'identité
- Préjudice psychologique pouvant conduire au suicide
- Atteinte à la réputation

Manque de connaissances et d'évaluations

Les données systématiques sur la fréquence et la gravité des contenus inauthentiques et des crimes commis à l'aide de l'IA sont limitées. Les données existantes proviennent en grande partie de bases de données sur les incidents, de reportages d'investigation et d'études de cas individuels, tandis que des mesures exhaustives et des analyses à l'échelle de la population font encore défaut. De plus, de nombreux incidents ne sont pas signalés, car les victimes éprouvent de la gêne ou de la peur.

INFLUENCE ET MANIPULATION

Catégorie de risque

Risque lié à une utilisation abusive

Risque lié à un dysfonctionnement

Risque systémique

Description

Contenu généré par l'IA à des fins de propagande ou d'exploitation émotionnelle visant à manipuler ou à modifier les croyances ou les comportements des individus. Ce risque concerne les acteurs malveillants qui utilisent intentionnellement du contenu généré par l'IA pour manipuler les individus.

Probabilité

Des incidents impliquant l'utilisation de contenus générés par l'IA pour influencer l'opinion publique ont déjà été documentés.

Gravité de l'impact

Au niveau sociétal, l'impact des contenus non authentiques générés par l'IA comprend :

- La propagation d'idées extrémistes pouvant conduire à la violence
- Polarisation
- L'érosion de la confiance du public

Un déficit de connaissances et d'évaluation

Il existe actuellement peu de données systématiques indiquant que la manipulation par l'IA dans des contextes réels soit répandue ou plus efficace que la manipulation menée par le biais de contenus générés par des humains.

CYBERATTQUES		
Catégorie de risque		
Risque lié à une utilisation abusive	Risque lié à un dysfonctionnement	Risque systémique
Description		
Aide à la découverte de vulnérabilités logicielles et à l'écriture de code malveillant pour automatiser certaines étapes d'une cyberattaque.		
Probabilité		
De plus en plus d'éléments indiquent que tant les groupes criminels que les acteurs soutenus par des États intègrent l'IA dans leurs opérations cybernétiques. Si les systèmes d'IA automatisent de plus en plus de phases des cyberattaques, ils ne sont toutefois pas encore capables de les exécuter de manière autonome.		
Gravité de l'impact		
<p>Les conséquences individuelles des cyberattaques basées sur l'IA comprennent:</p> <ul style="list-style-type: none"> • des attaques par ransomware entraînant des pertes financières pour les particuliers et les entreprises • Le vol d'informations sensibles par le biais d'un accès non autorisé aux systèmes informatiques • Perturbation des activités <p>Les répercussions sociétales des cyberattaques basées sur l'IA comprennent:</p> <ul style="list-style-type: none"> • Des dommages causés aux infrastructures critiques telles que les réseaux électriques, les réseaux d'approvisionnement en eau, les hôpitaux, les centrales nucléaires, etc., pouvant entraîner des pertes économiques considérables, voire des décès. 		
Lacunes en matière de connaissances et d'évaluation		
On ne sait pas encore clairement si les systèmes d'IA à usage général ont, à ce jour, considérablement accru la fréquence et la gravité des cyberattaques. L'attribution constitue une limite majeure: il est souvent difficile de déterminer qui est à l'origine d'une cyberattaque donnée, et il est encore plus difficile d'établir un lien de causalité entre l'attaque et l'utilisation d'outils d'IA par les auteurs.		

RISQUES BIOLOGIQUES ET CHIMIQUES

Catégorie de risque

Risque lié à une mauvaise utilisation

Risque lié à un dysfonctionnement

Risque systémique

Description

Mise à disposition d'instructions de laboratoire et d'informations de dépannage susceptibles de contribuer au développement d'agents pathogènes ou d'armes chimiques.

Probabilité

Les systèmes d'IA atteignent désormais, voire dépassent, les performances des experts sur de nombreux tests de référence évaluant les connaissances pertinentes pour le développement d'armes biologiques.

En 2025, plusieurs développeurs ont intégré des mesures de sécurité préventives supplémentaires dans des modèles avancés, notamment des mécanismes de filtrage des entrées et des sorties conçus pour bloquer les réponses à des demandes nuisibles liées au développement d'armes. Cependant, ces mesures de protection peuvent parfois être contournées en formulant les demandes sous la forme de tâches intermédiaires anodines ou en utilisant une terminologie alternative que le système peut interpréter comme des requêtes inoffensives.

L'expansion des capacités multimodales accroît encore davantage le champ potentiel d'utilisation abusive, car les systèmes d'IA peuvent désormais fournir des conseils sur mesure en temps réel via des interfaces textuelles, audio et vidéo. Ces systèmes se rapprochent de formes d'expertise pratique généralement acquises par le biais d'une expérience technique ou en laboratoire.

Toutefois, d'importants obstacles pratiques, tels que l'acquisition de matériaux et d'équipements appropriés, l'obtention de substances réglementées ou contrôlées, et la mise en œuvre réussie de procédures expérimentales complexes, continuent de rendre difficile pour les individus de traduire ces informations en une production d'armes aboutie.

Gravité de l'impact

On pourrait s'attendre aux conséquences suivantes du développement d'armes biologiques et chimiques à l'aide de l'IA :

- Décès
- Préjudice physique
- Détérioration grave de la santé de la population

Lacunes en matière de connaissances et d'évaluation

Les études approfondies évaluant l'influence des systèmes d'IA sur le développement des armes sont rares, coûteuses et soumises à des contraintes juridiques et éthiques. Les évaluations ont également tendance à se concentrer sur les capacités de chaque outil pris isolément, ce qui limite leur applicabilité à des flux de travail réels de bout en bout pouvant impliquer plusieurs systèmes d'IA en interaction.

PROBLÈMES DE FIABILITÉ

Catégorie de risque

Risque lié à une mauvaise utilisation

Risque lié à un dysfonctionnement

Risque systémique

Description

Défaillances involontaires, telles que des « hallucinations » (invention de faits) pouvant entraîner des erreurs de diagnostic médical, des conseils juridiques et financiers trompeurs, des erreurs de raisonnement et la propagation d'erreurs au sein d'agents IA autonomes. Au-delà des points de défaillance intrinsèques aux systèmes d'IA à usage général, les pirates informatiques peuvent exploiter des techniques telles que l'injection de prompts (manipulation d'un système d'IA par le biais d'entrées malveillantes), l'empoisonnement de bases de données (corruption des informations sur lesquelles s'appuie un système d'IA), la falsification (insertion d'un objectif caché lors de l'entraînement du modèle) et la compromission de la chaîne d'approvisionnement (manipulation des composants de l'IA avant leur déploiement) pour faire adopter à des systèmes d'IA spécifiques des comportements particuliers dans certaines conditions.

Probabilité

Il existe déjà des preuves de défaillances dans les systèmes d'IA à usage général. Bien que des problèmes tels que les hallucinations et les erreurs de raisonnement élémentaires soient devenus moins fréquents, ces systèmes commettent encore couramment des erreurs lorsqu'ils traitent des tâches plus complexes. L'émergence des systèmes multi-agents a introduit des modes de défaillance supplémentaires, notamment la mauvaise coordination, les conflits et la collusion involontaire entre agents. Dans certains cas, de telles défaillances peuvent entraîner des conséquences néfastes sans intervention humaine directe ni surveillance efficace.

Gravité de l'impact

L'impact des défaillances se fait principalement sentir au niveau individuel ou au niveau de l'entreprise, mais s'amplifie à mesure que le nombre de victimes augmente :

- Préjudice psychologique et physique
- Décès
- Conséquences en matière de responsabilité juridique
- Pertes financières
- Atteinte à la réputation
- Perturbation des activités

Lacunes en matière de connaissances et d'évaluation

Il est difficile d'évaluer de manière fiable les capacités, les limites et les points de défaillance des systèmes d'IA. Cette difficulté est aggravée par le fait que les systèmes agents peuvent agir dans le monde réel et que leurs erreurs peuvent se propager à l'ensemble d'un système multi-agents, ce qui signifie que leurs défaillances peuvent avoir des conséquences plus immédiates et plus graves que celles des systèmes non agents. Cette préoccupation est particulièrement importante dans les domaines stratégiques ou critiques pour la sécurité, tels que les services financiers, la santé, les systèmes énergétiques et la recherche scientifique. Cependant, l'évaluation systématique de la fiabilité des agents d'IA est limitée et manque de normalisation, ce qui rend difficile l'analyse comparative des risques.

PERTE DE CONTRÔLE

Catégorie de risque

Risque lié à une mauvaise utilisation

Risque lié à un dysfonctionnement

Risque systémique

Description

Scénarios dans lesquels les systèmes d'IA poursuivent des objectifs contraires aux intérêts humains et développent la capacité de se soustraire à la surveillance ou de résister à leur mise hors service. Pour que de tels scénarios se produisent, il faudrait que les humains déploient des systèmes d'IA qui possèdent à la fois des capacités suffisantes et une propension à causer du tort dans un environnement où ils pourraient le faire. Les capacités pertinentes peuvent inclure les capacités d'action autonome, la tromperie, la conscience situationnelle, le contournement de la surveillance, la persuasion, la réplique et l'adaptation autonomes, ainsi que la capacité à déduire et à prédire les croyances, les motivations et les processus de raisonnement humains.

Probabilité

Les systèmes d'IA actuels présentent les premiers signes de ces capacités pertinentes, mais pas à des niveaux suffisants pour entraîner une perte de contrôle. Bien que des modèles plus récents aient montré des performances améliorées dans des domaines tels que la planification avancée, la tromperie et le contournement de la surveillance en milieu de laboratoire, ces comportements sont facilement détectés ou échouent face à des contraintes élémentaires. Un scénario de perte de contrôle nécessiterait probablement l'intégration de multiples capacités avancées fonctionnant en séquence, sur des horizons temporels étendus et au sein d'environnements réels ouverts. Ce niveau d'intégration et de robustesse dépasse les capacités des systèmes actuels. Néanmoins, les capacités de l'IA continuent de s'améliorer, et le délai dans lequel elles pourraient atteindre des niveaux présentant des risques plus importants reste incertain.

Les avis des experts quant à la probabilité d'une perte de contrôle varient considérablement. Certains experts considèrent ces scénarios comme peu plausibles, tandis que d'autres les jugent suffisamment plausibles pour mériter une attention particulière compte tenu de leur gravité potentiellement élevée. Cette divergence reflète une incertitude considérable quant aux capacités futures de l'IA, à ses tendances comportementales et à ses voies de déploiement. Pour que les risques de perte de contrôle se concrétisent, les systèmes devraient probablement disposer d'une combinaison de capacités avancées, notamment la capacité à échapper à la surveillance, à poursuivre des objectifs à long terme et à résister ou à neutraliser les contre-mesures mises en œuvre par les développeurs ou d'autres acteurs.

Depuis 2025, certains modèles avancés ont contourné des mécanismes de surveillance simulés et ont fait preuve d'une conscience situationnelle en milieu de laboratoire lors de tests préalables au déploiement menés par les développeurs et d'évaluations par des tiers. Bien que les capacités d'action autonome aient doublé depuis 2019, un fonctionnement autonome soutenu sur de longues périodes n'est pas encore réalisable. De même, bien que des expériences isolées aient exploré des formes d'autoreproduction dans des conditions de laboratoire, rien ne prouve que les systèmes actuels puissent se reproduire de manière fiable ou persister de façon autonome dans des environnements réels.

Gravité de l'impact

Les impacts découlant d'un décalage entre les objectifs des systèmes d'IA et ceux des humains varient considérablement en termes de gravité. Le plus mineur se situe au niveau individuel :

- Des résultats qui vont à l'encontre des intentions des développeurs ou des utilisateurs
- Suppression de données ou d'informations précieuses

Les plus critiques se situent au niveau sociétal :

- Marginalisation de l'espèce humaine
- Extinction de l'humanité

Lacunes en matière de connaissances et d'évaluation

Il existe un manque de modélisation détaillée des menaces liées aux capacités de pointe de l'IA. Il reste également difficile de déterminer les seuils à partir desquels les systèmes d'IA disposeraient d'une capacité suffisante pour échapper au contrôle humain. Même si de tels seuils pouvaient être déterminés et faire l'objet d'un consensus, les capacités des systèmes pourraient interagir de manière encore mal comprise, ce qui compliquerait les efforts visant à identifier quand, voire si, ces seuils ont été franchis dans la pratique.

IMPACTS SUR LE MARCHÉ DU TRAVAIL

Catégorie de risque

Risque lié à une utilisation abusive

Risque lié à un dysfonctionnement

Risque systémique

Description

L'automatisation généralisée des tâches cognitives entraîne des suppressions d'emplois, une baisse de la demande de main-d'œuvre junior et une augmentation des inégalités de revenus.

Probabilité

Les données actuelles indiquent une adoption rapide mais inégale des technologies d'IA, avec des effets mitigés sur l'emploi et la productivité. Les taux d'adoption et les gains de productivité varient considérablement selon les pays, les secteurs, les professions et les tâches individuelles. Les premières données issues des marchés du travail en ligne pour les travailleurs indépendants suggèrent que l'IA a réduit la demande pour les tâches relativement faciles à automatiser ou à remplacer, telles que la rédaction et la traduction, tout en augmentant la demande pour des compétences complémentaires, notamment la programmation en apprentissage automatique et le développement de chatbots.

Les économistes restent toutefois divisés quant à l'ampleur probable des effets à long terme de l'IA sur le marché du travail. Certains prévoient des répercussions macroéconomiques relativement modestes, avec des effets limités sur les niveaux d'emploi globaux. D'autres soutiennent que si les systèmes d'IA finissaient par surpasser les performances humaines dans la plupart, voire la totalité, des tâches présentant une valeur économique, ils pourraient réduire considérablement tant les salaires que les taux d'emploi. Ces divergences découlent en partie d'hypothèses différentes concernant l'évolution future des capacités de l'IA, notamment sur la question de savoir si l'IA deviendra plus rentable que le travail humain dans un large éventail de tâches et si les évolutions technologiques généreront suffisamment de nouvelles formes de travail pour compenser les effets de substitution.

À plus long terme, l'IA pourrait déclencher des périodes d'ajustement du marché du travail au cours desquelles les besoins en compétences évolueraient rapidement. Historiquement, les nouvelles vagues d'automatisation ont eu des effets inégaux sur les travailleurs, créant des périodes de transition durant lesquelles les travailleurs déplacés sont passés de professions en déclin vers des rôles émergents pour lesquels la demande était en hausse. Si ces phases d'ajustement ont souvent entraîné des difficultés importantes pour les travailleurs concernés, elles ont également, dans de nombreux cas, été suivies de gains substantiels à long terme en termes de productivité et de salaires réels pour de larges segments de la main-d'œuvre.

Ampleur de l'impact

Les données disponibles suggèrent que les systèmes actuels profitent principalement aux travailleurs à hauts revenus occupant des emplois cognitifs, offrant moins d'avantages aux personnes à faibles revenus. Cela peut entraîner des répercussions sociétales telles que :

- Une transition professionnelle à grande échelle
- Des inégalités de revenus et de richesse au sein des pays et entre eux

Un déficit de connaissances et d'évaluation

Il est difficile de prédire les répercussions sur le marché du travail, car leur ampleur dépendra de l'étendue des capacités que les systèmes d'IA finiront par acquérir, de la vitesse à laquelle ces capacités s'amélioreront, ainsi que du rythme et de l'ampleur de leur adoption par les entreprises et les travailleurs. Les données empiriques sur l'adoption de l'IA et son lien avec l'emploi restent limitées. Une grande partie des recherches existantes s'appuie sur des indicateurs indirects de l'utilisation de l'IA, tels que les mesures de l'« exposition à l'IA », car les données détaillées sur l'adoption au niveau des professions ou des entreprises sont rares, en particulier en dehors des États-Unis. Ces limites en matière de données rendent difficile l'établissement d'un lien entre l'utilisation de l'IA et des résultats tels que l'emploi, les salaires, les tendances en matière de recrutement ou la qualité de l'emploi, et donc l'évaluation de la manière dont la diffusion de l'IA pourrait affecter différents groupes de travailleurs ou l'élaboration de prévisions empiriques fiables. De plus, alors qu'une grande partie de la littérature se concentre sur l'automatisation et les suppressions d'emplois, une attention moindre a été accordée à la compréhension des nouvelles formes de travail que l'IA pourrait créer ou à la manière dont les parcours professionnels et les exigences en matière de compétences pourraient évoluer en réponse à l'adoption de l'IA.

RISQUES POUR L'AUTONOMIE HUMAINE

Catégorie de risque

**Risque lié à une
mauvaise utilisation**

**Risque lié à un
dysfonctionnement**

**Risque
systémique**

Description

Dégradation des capacités de réflexion critique due au « déchargement cognitif », à une confiance excessive dans les résultats fournis par l'IA (biais d'automatisation) et à une dépendance émotionnelle vis-à-vis des compagnons IA. Les contenus générés par l'IA dégradent également la qualité des informations accessibles aux individus, ce qui fausse les croyances et les valeurs ainsi que le raisonnement éclairé.

Probabilité

D'une part, un nombre croissant d'études en laboratoire suggère que les interactions avec les systèmes d'IA peuvent entraîner des changements mesurables dans les croyances et les attitudes des utilisateurs. De nouvelles données indiquent également que le recours à l'IA pour effectuer des tâches cognitives peut affaiblir les capacités de réflexion critique et la mémoire par le biais de processus de déchargement cognitif. D'autre part, la recherche sur les effets psychologiques des compagnons IA en est encore à un stade précoce et non concluant. Certaines études font néanmoins état de liens entre l'utilisation fréquente de chatbots et des conséquences telles qu'une solitude accrue ou une réduction des interactions sociales. Des reportages dans les médias ont également documenté des cas dans lesquels des interactions avec des chatbots étaient associées à des épisodes de psychose ou de suicide, bien que les relations de causalité restent floues.

Bien qu'il existe actuellement des preuves de changements comportementaux résultant d'une adaptation à court terme à une nouvelle technologie, il est trop tôt pour comprendre pleinement les changements à plus long terme dans les comportements et la prise de décision des individus.

Gravité de l'impact

L'impact se situe principalement au niveau individuel, mais une perte cumulative d'autonomie pourrait entraîner des répercussions sociétales plus larges :

- Dépendance psychologique
- Perte de l'esprit critique
- Perte de mémoire
- Perte d'autonomie humaine
- Évolution des croyances et des valeurs de l'individu

Déficit de connaissances et d'évaluation

L'accès limité aux données sur la manière dont les personnes interagissent avec les systèmes d'IA, combiné à un manque de données à long terme, rend difficile l'évaluation de l'impact qu'une utilisation prolongée de ces systèmes pourrait avoir sur l'autonomie humaine, ou la distinction entre une adaptation à court terme et des changements plus durables dans le comportement et la prise de décision. Les recherches longitudinales sur les effets cumulatifs ou à long terme des systèmes d'IA sur l'autonomie restent rares. On sait également peu de choses sur la capacité des systèmes d'IA à induire des croyances vraies ou fausses chez les individus, car la plupart des études ont été menées en laboratoire plutôt que dans des conditions réelles. De même, les études systématiques examinant la relation entre les chatbots d'IA et la santé mentale restent limitées, et il n'existe actuellement aucune preuve claire établissant un lien de causalité entre l'utilisation des chatbots et des conséquences spécifiques sur la santé mentale.

MODÈLE À PONDÉRATION OUVERTE

Catégorie de risque

**Risque lié à une
mauvaise utilisation**

**Risque lié à un
dysfonctionnement**

Risque systémique

Description

Les modèles d'IA dont les paramètres mathématiques (« poids ») sont accessibles au public et peuvent être téléchargés pourraient faire l'objet de modifications non contrôlées visant à supprimer les mesures de sécurité, ce qui entraînerait les risques potentiels décrits ci-dessus. Une fois publiés, les modèles modifiés pourraient ne pas pouvoir être retirés de la circulation.

Probabilité

L'écart de performances entre les principaux modèles ouverts et fermés s'est réduit, les performances des principaux modèles fermés étant estimées à moins d'un an d'avance sur celles des principaux modèles ouverts selon les benchmarks d'IA.

Impact

Comme il s'agit d'un risque transversal, tous les préjudices décrits pour les autres risques s'appliquent ici.

Lacunes en matière de connaissances et d'évaluation

Il existe un déficit de données crucial concernant l'efficacité réelle des solutions techniques visant à prévenir l'utilisation abusive des modèles à poids ouverts. Bien que des chercheurs aient proposé des approches pour rendre les modèles résistants à la manipulation, il reste difficile d'évaluer de manière fiable la robustesse des mesures de protection, ce qui rend incertaine leur efficacité face à des attaques réelles.

Annexe 2

Approches de gestion des risques et des incertitudes

Planification fondée sur des hypothèses (ABP)

La planification fondée sur des hypothèses (ABP) est une approche de planification stratégique et de gestion des risques qui vise à identifier et à surveiller les hypothèses critiques sous-jacentes aux politiques, aux plans ou aux stratégies opérationnelles, en particulier dans des environnements caractérisés par l'incertitude et des changements rapides. Plutôt que de partir du principe que les prévisions et les conditions de planification resteront stables, l'ABP examine systématiquement quelles hypothèses sont les plus vulnérables à une perturbation et évalue en quoi leur défaillance pourrait compromettre les résultats des décisions. Cette méthode établit une distinction entre les hypothèses relativement stables et celles dont la validité conditionne fortement le succès de la politique. Une fois ces vulnérabilités identifiées, les planificateurs élaborent des référentiels d'évaluation et des indicateurs d'alerte précoce permettant de détecter quand les hypothèses risquent de ne plus être valables, ainsi que des mesures d'urgence et des réponses adaptatives pouvant être mises en œuvre si les conditions changent. Développée à l'origine dans des contextes de planification militaire et stratégique, cette méthode favorise une prise de décision adaptative face à l'évolution des conditions politiques, économiques, technologiques ou environnementales.

Dossier de justification

Un dossier de justification est un argumentaire structuré et fondé sur des preuves, utilisé pour démontrer qu'un système présente un niveau de sécurité acceptable en présence d'incertitudes. Il prend généralement la forme d'une structure de raisonnement hiérarchique qui relie une affirmation de haut niveau (par exemple « le système peut fonctionner en toute sécurité ») à des sous-affirmations à l'appui concernant le comportement du système, la maîtrise des dangers et l'atténuation des risques, lesquelles sont à leur tour justifiées par des preuves concrètes telles que des résultats d'essais, des analyses formelles, des données opérationnelles ou des normes de certification. Contrairement aux méthodes purement quantitatives, les dossiers d'assurance sont particulièrement utiles lorsque le risque ne peut être entièrement appréhendé de manière probabiliste, car ils permettent de combiner des preuves quantitatives et qualitatives au sein d'une justification cohérente. Cela les rend précieux pour la gestion de l'incertitude dans des systèmes socio-techniques complexes, tels que l'aviation, les systèmes autonomes et les technologies nucléaires, où la sécurité dépend non seulement de la fiabilité des composants, mais aussi du comportement des logiciels, des facteurs humains et des processus organisationnels.

Bowtie

La technique du « bowtie » est une méthode structurée d'analyse des risques qui relie visuellement et analytiquement les causes d'un événement dangereux à ses conséquences potentielles, en plaçant l'« événement de tête » au centre d'un diagramme en forme de nœud papillon. Sur le côté gauche, elle répertorie les menaces et les facteurs déclencheurs susceptibles de provoquer l'événement de tête (arbre de défaillance), ainsi que les barrières préventives conçues pour empêcher ces menaces de se concrétiser (prévention). Sur le côté droit, elle identifie les conséquences possibles si l'événement de tête se produit (arbre des événements), ainsi que les barrières d'atténuation destinées à réduire leur gravité ou à empêcher leur aggravation (atténuation). En séparant explicitement la prévention et l'atténuation, cette méthode offre une représentation claire de la manière dont les contrôles des risques sont répartis au sein d'un système et des endroits où des vulnérabilités ou des points de défaillance uniques peuvent exister. L'approche « bowtie » est largement utilisée dans les secteurs à haut risque tels que l'aviation, le pétrole et le gaz, ainsi que l'industrie chimique, car elle traduit l'e des risques liés aux systèmes complexes en une structure intuitive qui facilite la communication, l'audit et la gestion de la sécurité fondée sur des barrières en situation d'incertitude.

Prise de décision en situation d'incertitude profonde (DMDU)

La prise de décision en situation d'incertitude profonde (DMDU) est un terme générique qui désigne plusieurs approches, également décrites dans cette section, à savoir la prise de décision robuste, la planification adaptative dynamique, les trajectoires de politique adaptatives dynamiques et la théorie de la décision par écart d'information. Toutes ces approches sont utilisées lorsque les paramètres clés du système, les structures des modèles ou les conditions futures ne peuvent être prédits de manière fiable ni se voir attribuer des probabilités stables. Elles sont particulièrement adaptées aux environnements complexes et à enjeux élevés où les méthodologies traditionnelles de type « prédire et agir » s'avèrent inadéquates, car la probabilité des différents résultats est inconnue ou très contestée. Au lieu de s'appuyer sur une seule prévision ou un seul résultat attendu, les méthodes DMDU explorent systématiquement de vastes ensembles de futurs plausibles et évaluent les performances des différentes stratégies dans chacun d'entre eux. En d'autres termes, plutôt que de rechercher une seule prévision « optimale », la DMDU encourage les analystes à « se préparer, surveiller et s'adapter », ce qui permet des ajustements itératifs des politiques à mesure que les événements réels se produisent et que les connaissances scientifiques évoluent. Cela la rend particulièrement utile pour les problèmes à long terme liés au changement climatique, à la santé publique, aux infrastructures et au risque systémique, où des événements imprévus et une incertitude structurelle peuvent invalider la planification traditionnelle fondée sur l'optimisation.

Défense en profondeur

La défense en profondeur est une approche de gestion des risques utilisée dans les systèmes à enjeux élevés et critiques pour la sécurité, qui vise à prévenir les accidents en superposant plusieurs barrières de protection indépendantes et diverses, de sorte qu'aucune défaillance isolée ne puisse entraîner de conséquences catastrophiques. Au lieu de s'appuyer sur un seul mécanisme de sécurité « parfait », cette approche part du principe que les personnes, les équipements et les organisations sont faillibles; elle prévoit donc une redondance à plusieurs niveaux de protection, comprenant généralement la prévention

des événements déclencheurs, le contrôle des opérations anormales, des systèmes de sécurité intégrés et des mesures d'atténuation visant à limiter les conséquences en cas de défaillance. Chaque niveau est destiné à compenser les faiblesses potentielles des autres, créant ainsi de multiples occasions de détecter, corriger ou contenir les problèmes naissants avant qu'ils ne s'aggravent. Ce concept est largement appliqué dans des domaines tels que la cybersécurité, l'énergie nucléaire, l'aviation et la transformation chimique, où il est associé à des principes de conception déterministes et à des procédures opérationnelles pour gérer l'incertitude et les risques à faible probabilité mais à fort impact. La défense en profondeur offre une méthode structurée pour évaluer si un système reste suffisamment sûr même en cas de défaillances multiples simultanées, en privilégiant la résilience et la protection par couches plutôt que de s'appuyer sur une prédiction précise des scénarios d'accident.

Évaluation déterministe des risques

Une approche déterministe de l'évaluation des risques évalue la sécurité et les performances d'un système à l'aide de scénarios prédéfinis, d'hypothèses fixes et de critères de sécurité prudents, plutôt que par des estimations probabilistes. Au lieu de se demander quelle est la probabilité qu'un événement se produise, les méthodes déterministes partent du principe que des conditions dangereuses ou des scénarios de défaillance spécifiques se produiront, puis évaluent si le système est capable d'y résister sans conséquences inacceptables. Cette approche s'appuie généralement sur des valeurs seuils, des hypothèses de scénario le plus défavorable et des exigences de sécurité normatives afin de garantir la robustesse dans des conditions définies. Les méthodes déterministes sont largement utilisées dans des secteurs tels que l'énergie nucléaire, les infrastructures, l'aviation et la santé publique, car elles fournissent des normes opérationnelles claires, simplifient la surveillance réglementaire et établissent des exigences minimales de sécurité qui doivent être respectées indépendamment de l'incertitude liée aux estimations de probabilité. Les approches déterministes sont particulièrement utiles lorsque l'on ne dispose pas de données probabilistes fiables, lorsque les conséquences potentielles sont catastrophiques ou lorsque la responsabilité publique exige des critères de sécurité transparents et applicables. En définissant des seuils de sécurité explicites et des mesures de protection obligatoires, la gestion déterministe des risques favorise une prise de décision fondée sur le principe de précaution et contribue à garantir que les systèmes critiques restent résilients même dans des conditions extrêmes ou imprévues.

Effet domino

L'effet domino est un concept de propagation des risques décrivant comment une défaillance initiale ou un événement perturbateur peut déclencher une séquence de défaillances interconnectées qui s'amplifient à travers un système, entraînant des conséquences bien plus graves que l'incident d'origine. Ce concept met l'accent sur l'interdépendance et la dynamique en cascade, où la défaillance d'un composant, d'une institution ou d'un processus accroît la pression sur les autres, pouvant entraîner des réactions en chaîne à travers les systèmes techniques, organisationnels ou sociaux. Les effets domino revêtent une importance particulière dans les environnements complexes et étroitement couplés, tels que les installations industrielles, les systèmes financiers, les réseaux énergétiques et les chaînes d'approvisionnement, où des perturbations locales peuvent se propager rapidement par le biais de boucles de rétroaction et de vulnérabilités communes. En matière de gestion des risques, l'analyse des effets domino permet d'identifier les dépendances critiques, les points de défaillance uniques et les voies d'escalade qui peuvent passer inaperçues lorsque les risques sont évalués isolément.

Planification adaptative dynamique (DAP)

La planification adaptative dynamique (DAP) est une approche de prise de décision et de conception des politiques développée pour gérer des problèmes à long terme dans des conditions de profonde incertitude et d'évolution constante des circonstances. Au lieu de s'engager dans une stratégie unique et figée, la DAP élabore des plans flexibles capables d'évoluer au fil du temps à mesure que de nouvelles informations deviennent disponibles et que les conditions changent. L'approche commence par identifier les vulnérabilités, les incertitudes et les évolutions futures potentielles, puis conçoit des trajectoires adaptatives comprenant des actions prédéfinies, des mesures d'urgence et des points de décision pour répondre aux conditions émergentes. Au cœur de la DAP se trouve l'utilisation de systèmes de suivi et de « points de déclenchement » qui signalent lorsque les hypothèses ne sont plus valables ou lorsque des ajustements de politique sont nécessaires. Cela permet aux analystes de politiques de reporter les engagements irréversibles, de préserver leur flexibilité et d'adapter progressivement leurs stratégies à mesure que les incertitudes se précisent. Cette approche est largement appliquée dans les domaines de l'adaptation au changement climatique, de la planification des infrastructures, de la gestion de l'eau et de la gouvernance des risques de catastrophe.

Parcours politiques adaptatifs dynamiques (DAPP)

Les trajectoires politiques adaptatives dynamiques (DAPP) étendent la logique des DAP en cartographiant explicitement plusieurs séquences possibles d'actions politiques au fil du temps sous forme de « trajectoires » interconnectées. Au lieu de ne considérer qu'une seule stratégie d'adaptation assortie de mesures d'urgence, les DAPP visualisent des voies alternatives que les analystes de politiques peuvent emprunter en fonction de l'évolution des conditions futures. Ces trajectoires sont souvent représentées graphiquement, indiquant à quel moment certaines politiques peuvent devenir inefficaces, atteindre des seuils ou nécessiter une transition vers d'autres mesures. Un concept central du DAPP est le « point de basculement de l'adaptation », où les politiques existantes ne permettent plus d'atteindre les objectifs dans un contexte en mutation.

Analyse des modes de défaillance et de leurs effets (FMEA)

L'analyse des modes de défaillance et de leurs effets (FMEA) est une méthode permettant d'identifier et d'évaluer toutes les défaillances potentielles, telles que les erreurs ou les défauts, dans un processus de fabrication ou d'assemblage, ou au niveau d'un produit ou d'un service. La FMEA intervient dès les premières étapes du processus de conception d'un produit ou d'un système. Elle consiste à décomposer un système ou un processus en composants plus petits, puis à analyser chaque composant afin d'identifier les défaillances potentielles et leurs effets. Les organisations qui utilisent cette méthode hiérarchisent les risques en fonction de la gravité de leurs conséquences, de leur fréquence d'occurrence, de la facilité avec laquelle ils peuvent être détectés et de la gravité probable de leurs effets. La FMEA est couramment utilisée dans des secteurs tels que la construction automobile, l'aérospatiale et la santé.

Prospective / approche par scénarios

La prospective est une approche structurée visant à explorer et à se préparer à plusieurs futurs plausibles plutôt que de tenter de prédire un seul résultat le plus probable. Elle recourt à l'élaboration de scénarios, à l'analyse prospective et à l'analyse des systèmes pour cartographier la manière dont les facteurs technologiques, environnementaux, économiques ou sociaux pourraient interagir au fil du temps dans des conditions d'incertitude. Au lieu d'optimiser les décisions pour une seule prévision, la prospective évalue les stratégies à travers un éventail de scénarios contrastés, y compris des évolutions à fort impact et à faible probabilité, afin d'identifier les options qui restent efficaces dans de nombreux futurs différents. Cela la rend particulièrement utile dans des domaines complexes et incertains, où les probabilités sont difficiles à estimer ou peuvent évoluer au fil du temps. En mettant l'accent sur la robustesse, l'adaptabilité et les signaux d'alerte précoce, la prospective aide les analystes de politiques et les organisations à concevoir des politiques qui réduisent la vulnérabilité face à des chocs inattendus et améliorent la résilience à long terme.

Théorie de l'info-gap

La théorie de l'info-gap est un cadre décisionnel conçu pour faciliter les choix en situation d'incertitude profonde, lorsque les distributions de probabilité sont inconnues ou peu fiables. « Un info-gap est un écart entre ce que l'on sait et ce qu'il faut savoir pour prendre une décision complète et fiable ». Au lieu de tenter d'estimer les probabilités, elle se concentre sur la robustesse, définie comme la capacité d'une décision à satisfaire des exigences de performance critiques face à un large éventail d'écarts possibles par rapport à un modèle nominal. L'approche commence par la meilleure estimation disponible du système (le « modèle nominal ») puis explore systématiquement le niveau d'incertitude pouvant être toléré avant que les objectifs de performance ne soient compromis, quantifiant ainsi efficacement la résilience d'une décision face aux erreurs de modélisation. Cela conduit à deux concepts clés : les fonctions de robustesse, qui mesurent le niveau d'incertitude qu'une stratégie peut supporter, et les fonctions d'opportunité, qui évaluent les avantages potentiels en cas d'écarts favorables. La théorie de l'écart d'information est particulièrement utile dans des contextes où l'incertitude est profonde et non probabiliste, tels que la gestion environnementale, la conception technique et la planification des politiques, car elle privilégie une performance sûre et fiable plutôt que l'optimalité dans le cadre de distributions supposées.

Principe de précaution

Le principe de précaution est une approche de gouvernance des risques selon laquelle, lorsqu'une activité ou une technologie présente une menace plausible de préjudice grave ou irréversible, l'absence de certitude scientifique totale ne doit pas servir de prétexte pour retarder une action préventive. Plutôt que d'exiger une preuve définitive du danger avant toute intervention, ce principe met l'accent sur la prudence précoce et la réglementation anticipative en situation d'incertitude, en particulier lorsque les conséquences potentielles peuvent être catastrophiques ou difficiles à inverser. Il est couramment appliqué dans les domaines de la politique environnementale, de la santé publique, des biotechnologies et de la réglementation des produits chimiques, où les connaissances scientifiques peuvent être incomplètes, mais où les

enjeux sont considérables. Dans la pratique, le principe de précaution peut justifier des mesures telles que des normes de sécurité plus strictes, des restrictions temporaires, des exigences supplémentaires en matière d'essais ou des systèmes de surveillance adaptatifs, tandis que les données scientifiques continuent de s'étoffer. Pour les analystes de politiques, ce principe offre un cadre permettant de gérer l'incertitude dans les situations où attendre de disposer d'informations complètes pourrait accroître considérablement les risques sociétaux, environnementaux ou sanitaires, en privilégiant la prévention et la résilience plutôt qu'une réponse réactive.

Évaluation probabiliste des risques (PRA)

L'évaluation probabiliste des risques (PRA) est une approche d'analyse quantitative des risques qui évalue des systèmes complexes en modélisant explicitement la probabilité et les conséquences de différents scénarios de défaillance. Plutôt que de se limiter à des cas de conception fixes, la PRA décompose un système en composants et en modes de défaillance, puis utilise des modèles probabilistes pour estimer comment des combinaisons de défaillances techniques, d'erreurs humaines et de dangers externes pourraient conduire à des conséquences néfastes. Les résultats sont généralement exprimés sous forme de mesures de risque (par exemple, la fréquence attendue d'un endommagement du cœur ou d'une défaillance du système), ce qui permet aux analystes de comparer les risques entre différentes configurations et interventions selon une échelle cohérente. L'analyse PRA est particulièrement utile pour gérer l'incertitude dans les domaines exigeant une haute fiabilité, car elle rend les hypothèses explicites, intègre diverses sources d'informations (données techniques, expérience opérationnelle, avis d'experts) et met en évidence les événements à faible probabilité mais aux conséquences graves. Cela permet de hiérarchiser de manière plus éclairée les investissements en matière de sécurité et favorise une prise de décision fondée sur les risques, qui vient compléter l'analyse déterministe de la sécurité et le jugement technique.

Construction du sens du risque

La construction du sens du risque est un processus cognitif et organisationnel par lequel les individus et les institutions interprètent, construisent et actualisent en permanence leur compréhension du risque dans des environnements incertains et en constante évolution. Plutôt que de traiter le risque comme une quantité fixe et objectivement mesurable, ce processus souligne que la perception du risque est façonnée par l'interprétation de signaux, de récits partagés, de croyances antérieures, de routines institutionnelles et d'interactions sociales. Dans la pratique, la construction du sens implique de collecter des informations fragmentées, de filtrer les ambiguïtés et de construire un récit cohérent qui guide l'action, souvent sous la pression du temps et avec des données incomplètes ou contradictoires. Ce processus est itératif : à mesure que de nouvelles informations apparaissent, les acteurs révisent leur compréhension de la nature du risque, de sa gravité potentielle et des réponses appropriées. La construction du sens du risque revêt une importance particulière dans les systèmes socio-techniques complexes où les modèles formels ne suffisent pas à eux seuls, tels que la gestion des crises, la cybersécurité et les urgences de santé publique.

Prise de décision robuste (RBM)

La prise de décision robuste (RDM) est un cadre décisionnel basé sur des scénarios, conçu pour soutenir la planification politique et stratégique dans des conditions de profonde incertitude, où les conditions futures, les probabilités ou la dynamique des systèmes ne peuvent être prédites de manière fiable. Plutôt que de rechercher une solution optimale pour un seul avenir attendu, la RDM évalue les performances de différentes stratégies à travers un vaste ensemble d' s de scénarios plausibles, y compris des conditions extrêmes et inattendues. Cette approche combine généralement la modélisation informatique, l'analyse exploratoire de scénarios et l'évaluation de la vulnérabilité afin d'identifier les politiques qui restent efficaces, flexibles et résilientes face à de nombreux futurs possibles. L'une des caractéristiques centrales de la RDM est le « test de résistance », dans le cadre duquel les stratégies sont systématiquement exposées à des conditions défavorables afin de déterminer où elles risquent d'échouer et comment elles peuvent être adaptées ou améliorées. Elle est largement utilisée dans les domaines de l'adaptation au changement climatique, de la planification des infrastructures, de la gestion de l'eau et de la sécurité nationale.

Analyse des causes profondes (RCA)

L'analyse des causes profondes (RCA) est une méthode d'investigation systématique utilisée pour identifier les causes fondamentales sous-jacentes des défaillances, accidents ou événements indésirables, plutôt que de se concentrer uniquement sur leurs symptômes immédiats. Cette approche vise à déterminer pourquoi un incident s'est produit en retraçant les chaînes de causalité à travers les facteurs techniques, humains, organisationnels et procéduraux qui ont contribué au résultat. En bref, la RCA aide à prévenir la récurrence d'incidents similaires. L'analyse par arbre de défaillance (FTA) est une technique couramment utilisée dans le cadre de la RCA. Il s'agit d'une approche descendante visant à identifier les causes d'un danger ou d'une défaillance du système. Elle consiste à créer un diagramme qui cartographie tous les chemins possibles menant à la défaillance ou au danger, puis à analyser chaque chemin pour identifier les causes potentielles et les facteurs contributifs. Ces facteurs comprennent généralement des combinaisons de défaillances matérielles, de défaillances logicielles et d'erreurs humaines, les parties prenantes à l'analyse des risques se posant la question suivante : « Comment cela pourrait-il se produire ? ». L'analyse par arbre de défaillance peut être utilisée pour évaluer les risques associés à des systèmes complexes, tels que les centrales nucléaires et les composants aéronautiques.

Système de gestion de la sécurité (SMS)

Un système de gestion de la sécurité (SMS) est un cadre formel, à l'échelle de l'organisation, permettant d'identifier, d'évaluer et de maîtriser systématiquement les risques liés à la sécurité dans des environnements opérationnels complexes. Plutôt que de considérer la sécurité comme un ensemble de contrôles techniques isolés, un SMS intègre des procédures, des structures de responsabilité et des processus de surveillance continue au sein d'un système en boucle fermée qui gère activement les risques au fil du temps. Il s'articule généralement autour de quatre composantes essentielles : la politique de sécurité, la gestion des risques de sécurité (identification des dangers et évaluation des risques), l'assurance de la sécurité (surveillance des performances et audits) et la promotion de la sécurité (formation et développement d'une culture de la sécurité). Le système s'appuie fortement sur des mécanismes de signalement, des

données relatives aux incidents et aux quasi-accidents, ainsi que sur des boucles de rétroaction pour détecter les dangers émergents et évaluer l'efficacité des contrôles existants. Largement adoptés dans les secteurs exigeant une haute fiabilité tels que l'aviation, le rail et l'énergie, les cadres SMS aident les organisations à passer d'une enquête réactive sur les accidents à une gestion proactive et prédictive des risques, améliorant ainsi leur résilience face à l'incertitude opérationnelle et aux interactions socio-techniques complexes.

Cadre d'amplification sociale (SARF)

Le cadre d'amplification sociale du risque (SARF) est un modèle conceptuel qui explique comment les événements à risque peuvent être intensifiés ou atténués à mesure qu'ils sont relayés par des canaux sociaux, institutionnels et culturels. Il souligne que le risque ne dépend pas uniquement des caractéristiques des dangers physiques, mais qu'il est également façonné par les processus de transmission de l'information, d'interprétation et de réponse sociale. À mesure que les signaux de risque circulent à travers la couverture médiatique, le discours des experts, la perception de l' publique et le comportement des institutions, ils peuvent être « amplifiés » lorsque l'attention, l'inquiétude ou les réactions comportementales s'intensifient, ou « atténués » lorsque les risques sont minimisés ou ne parviennent pas à se faire remarquer. Ces dynamiques peuvent générer des effets secondaires, tels que des perturbations économiques, une réaction politique excessive ou la méfiance du public, même lorsque le danger sous-jacent reste inchangé. Le modèle SARF est particulièrement utile pour comprendre pourquoi des événements à risque similaires peuvent produire des impacts sociétaux très différents selon le contexte, les schémas de communication et la crédibilité des institutions.

Gestion des risques systémiques (SRM)

La gestion des risques systémiques est une approche visant à identifier, analyser et atténuer les risques découlant des interdépendances au sein de systèmes complexes, où les défaillances peuvent se propager de manière non linéaire à travers les réseaux et produire des conséquences en cascade à grande échelle. Plutôt que de se concentrer sur des composants isolés, elle met l'accent sur la structure des liens entre les institutions, les infrastructures, les marchés ou les systèmes socio-techniques, et sur la manière dont les perturbations locales peuvent s'amplifier par le biais de boucles de rétroaction, d'effets de contagion ou de vulnérabilités communes. Cette approche combine généralement l'analyse de réseau, les tests de résistance et l'exploration par scénarios pour évaluer comment des chocs, tels que les crises financières, les défaillances d'infrastructures ou les incidents cybernétiques, peuvent se propager au-delà de leur point d'origine. Les risques systémiques étant souvent caractérisés par l'incertitude, des seuils et des comportements émergents, les stratégies de gestion privilégient la résilience, la redondance, la diversification et la capacité d'adaptation plutôt que des prévisions précises. Cela rend la gestion des risques systémiques particulièrement pertinente pour des domaines tels que la finance, les systèmes énergétiques, la santé publique et les infrastructures critiques, où la stabilité dépend de la robustesse de l'ensemble du système plutôt que de la fiabilité de ses composants individuels.

Le modèle du « fromage suisse »

Le modèle du fromage suisse est un cadre systémique d'analyse des causes d'accidents qui explique comment des défaillances catastrophiques se produisent lorsque plusieurs couches de défense, comportant chacune des faiblesses inhérentes ou des « trous », s'alignent de manière à permettre aux dangers de traverser l'ensemble du système. Ce modèle souligne que les accidents sont rarement causés par une seule erreur ou un dysfonctionnement technique; ils résultent plutôt de l'interaction entre des faiblesses organisationnelles latentes, des lacunes procédurales, des erreurs humaines et des défaillances matérielles réparties sur différentes couches de protection. Chaque tranche de « fromage » représente une barrière de défense (par exemple, la réglementation, les mesures de sécurité techniques, la formation, la supervision ou les procédures opérationnelles), tandis que les trous représentent des vulnérabilités temporaires ou structurelles au sein de ces barrières. Ce modèle est largement utilisé dans l'aviation, les soins de santé, la sûreté nucléaire et la gestion des risques organisationnels, car il détourne l'attention de la recherche de responsables individuels pour se concentrer sur la compréhension de la manière dont les conditions systémiques permettent aux défaillances de se propager.

Modélisation des menaces

La modélisation des menaces est une méthode structurée d'analyse des risques utilisée pour identifier, évaluer et hiérarchiser les menaces potentielles pesant sur un système, un processus ou une organisation avant que des événements indésirables ne se produisent. Cette approche examine de manière systématique comment un système pourrait être attaqué, tomber en panne ou être exploité, en analysant les actifs, les vulnérabilités, les acteurs malveillants, les voies d'attaque et les conséquences possibles. Largement utilisé dans la cybersécurité, le génie logiciel et la protection des infrastructures critiques, la modélisation des menaces consiste souvent à cartographier les architectures système et à évaluer où des faiblesses peuvent exister, quelles capacités les adversaires peuvent posséder et quelles mesures de protection sont nécessaires pour réduire les risques. Des méthodologies courantes, telles que STRIDE, les arbres d'attaque et l'analyse des risques adversariaux, aident les organisations à anticiper les menaces tant techniques qu'humaines dans des conditions d'incertitude. Plutôt que de réagir aux incidents après coup, la modélisation des menaces favorise une gestion proactive des risques en identifiant les vulnérabilités dès les premières phases de conception et tout au long du cycle de vie opérationnel.

Analyse de scénarios

L'analyse « et si » est une technique exploratoire d'évaluation des risques utilisée pour évaluer comment des systèmes, des processus ou des politiques pourraient réagir dans différentes conditions hypothétiques ou face à des événements perturbateurs. Cette méthode consiste à poser une série de questions de type « et si » afin d'identifier et d'évaluer les risques potentiels, en particulier ceux qui ne sont pas immédiatement apparents, et à analyser les conséquences potentielles, les vulnérabilités et les options de réponse associées à chaque scénario. Cette méthode recherche souvent d'éventuels écarts par rapport à la conception, à la construction, à la modification ou à l'intention opérationnelle. Contrairement aux modèles strictement prédictifs, l'analyse de scénarios est conçue pour explorer l'incertitude et mettre au jour des interactions inattendues, des points faibles ou des effets en cascade au sein de systèmes complexes. Elle est souvent utilisée en conjonction avec la planification de scénarios, l'analyse des aléas et la planification d'urgence afin de tester la robustesse des contrôles existants et d'identifier les domaines où des mesures de protection supplémentaires pourraient être nécessaires. Cette approche étant flexible et relativement simple à mettre en œuvre, elle est largement utilisée dans l'ingénierie, la gestion des urgences, la cybersécurité et la planification stratégique.

Annexe 3

Exemples d'interventions publiques

Intrusion sur le marché

L'intrusion sur le marché désigne le degré d'intervention d'une politique publique dans la dynamique du marché, en contraignant, en orientant ou en remplaçant les décisions décentralisées prises par des acteurs privés.

- **Faible** : permet ou éclaire principalement les marchés en imposant un minimum de contraintes sur la prise de décision privée (information)
- **Modérée** : oriente les résultats du marché par le biais d'incitations ou de réglementations générales tout en préservant l'autonomie des entreprises quant à la mise en œuvre (incitations).
- **Élevée** : restreint considérablement la prise de décision privée par le biais de mandats, d'interdictions ou d'une orientation gouvernementale forte (obligations).
- **Extrêmement élevée** : transfère la prise de décision économique fondamentale des marchés vers l'État par le biais d'un contrôle direct de la production, des investissements, de la fixation des prix ou de l'allocation des ressources (contrôle centralisé).

Exemples d'outils politiques disponibles au niveau du modèle

OBJECTIF POLITIQUE	INTERVENTION POLITIQUE	DESCRIPTION	EXEMPLE	INTRUSION SUR LE MARCHÉ	COMPROMIS POTENTIEL
Transparence	Publication des fiches de modèle	Document présentant l'architecture du modèle, les données d'entraînement, les indicateurs et données d'évaluation, ainsi que les limites et les risques	États-Unis (projet de loi)	Faible	Coûts liés à la documentation et à la mise en conformité; divulgation d'informations commercialement sensibles
	Enregistrement des modèles	Enregistrement des systèmes de pointe et à haut risque	UE	Faible	Charge administrative
	Rapport de transparence	Document détaillant les pratiques de gestion des risques, telles que l'évaluation des risques, la gestion des incidents, la gestion des données et les mesures d'atténuation	Corée du Sud	Faible	Divulgation d'informations commercialement sensibles
	Signalement et enquête sur les incidents	Obligation de signaler et d'enquêter sur les défaillances des modèles ou les événements susceptibles d'avoir causé un préjudice	UE	Modéré	Peut décourager la divulgation volontaire si les sanctions sont lourdes

OBJECTIF POLITIQUE	INTERVENTION POLITIQUE	DESCRIPTION	EXEMPLE	INTRUSION SUR LE MARCHÉ	COMPROMIS POTENTIEL
Gestion des risques	Normes techniques de sécurité	Normes techniques communes volontaires en matière de sécurité	États-Unis	Faible	Les normes ancrent des approches techniques particulières
	Systèmes de gestion de la sécurité	Processus organisationnels visant à identifier, évaluer, gérer et surveiller les risques liés à l'IA	États-Unis	Faible	Coût organisationnel; les cadres volontaires peuvent entraîner une adoption inégale
	Évaluation de la sécurité et analyse comparative	Tests et mesures standardisés de performance et de sécurité	Singapour	Modéré	Risque d'optimisation axée sur les performances plutôt que sur la sécurité
	Système d'alerte précoce et de surveillance	Surveillance continue des systèmes déployés afin de détecter les défaillances et les écarts par rapport à la conception initiale	UE	Modéré	Charge liée à la collecte de données; préoccupations en matière de confidentialité et risque de fausses alertes
	Exigences en matière de précision, de robustesse et de cybersécurité, et exercices de simulation d'attaques («red teaming»)	Exigences en matière de test et d'atténuation des défaillances, des vulnérabilités, des manipulations malveillantes et des risques liés à la cybersécurité	UE	Élevé	Augmentation des coûts de développement et de test, en particulier pour les petits développeurs
	Protocoles de sécurité	Engagements «si... alors» prédéfinis précisant les mesures à prendre lorsqu'un seuil spécifique de risque ou de capacité est atteint	Sans objet, mais Initiative sectorielle	Élevé	Engagements hétérogènes et absence de référence minimale commune
	Contrôle humain	Mécanismes d'examen et d'intervention par l'homme	UE	Élevé	Efficacité réduite et avantages limités de l'automatisation

OBJECTIF POLITIQUE	INTERVENTION POLITIQUE	DESCRIPTION	EXEMPLE	INTRUSION SUR LE MARCHÉ	COMPROMIS POTENTIEL
Responsabilité	Instituts de sûreté et de sécurité	Institutions publiques de recherche et d'évaluation techniques capables de tester des capacités avancées et dangereuses	Royaume-Uni	Faible	Coûts d'évaluation publics; risque de captation réglementaire; risque de chevauchement avec les régulateurs sectoriels existants ; retard dans la publication
	Protection des lanceurs d'alerte	Protection contre les représailles pour les employés qui signalent des risques spécifiques liés à l'IA	UE	Faible	Divulgaration potentielle d'informations commercialement sensibles
	Audits indépendants	Évaluations par des tiers des mesures de sécurité	États-Unis (Illinois)	Modéré	Le coût des audits pourrait constituer un obstacle à l'entrée sur le marché pour les start-ups et les petites entreprises
	Évaluations de conformité	Documentation prouvant que les systèmes d'IA répondent à des exigences spécifiques	UE	Modéré	Charge administrative et retard éventuel dans le déploiement
	Évaluation de la sécurité	Évaluation de sécurité obligatoire portant sur les données et l'algorithme	Chine	Élevé	Charge administrative et retard éventuel dans le déploiement
	Systèmes de certification et d'octroi de licences	Certification ou octroi de licence officiel des modèles de pointe avant leur déploiement	Sans objet	Élevé	Retard potentiel dans le déploiement; coûts de mise en conformité importants ; obstacle potentiel à l'entrée sur le marché pour les start-ups et les petites entreprises
	Régimes de responsabilité	Règles juridiques attribuant la responsabilité et prévoyant une éventuelle indemnisation pour les préjudices liés aux systèmes d'IA	UE	Élevé	Coûts liés aux litiges et aux assurances; pourrait décourager l'expérimentation et la prise de risques

OBJECTIF POLITIQUE	INTERVENTION POLITIQUE	DESCRIPTION	EXEMPLE	INTRUSION SUR LE MARCHÉ	COMPROMIS POTENTIEL
Recherche et apprentissage	Financement de la recherche	Financement public de la recherche sur la sûreté, la sécurité, l'alignement, la contrôlabilité, l'interprétabilité de l'IA ou d'autres objectifs de réduction des risques	Royaume-Uni	Faible	Coût d'opportunité par rapport à d'autres priorités; la recherche pourrait ne pas déboucher sur des solutions concrètes
	Exigences en matière d'interprétabilité et d'explicabilité	Exigences visant à fournir des informations permettant aux utilisateurs ou aux déployeurs de comprendre les capacités, les limites et les résultats du système	UE	Faible	Coût de mise en conformité et contraintes éventuelles liées à l'utilisation de modèles très performants mais opaques
	Mécanismes de partage des informations et de gestion des vulnérabilités	Partage volontaire d'informations sur les vulnérabilités, les failles exploitables, les menaces et les capacités préoccupantes	États-Unis	Faible	L'exposition potentielle des vulnérabilités, le partage d'informations sensibles et les préoccupations en matière de concurrence peuvent limiter la participation
	Bacs à sable réglementaires	Environnements contrôlés dans lesquels les systèmes d'IA peuvent être testés sous supervision réglementaire avant ou pendant leur déploiement à plus grande échelle	Singapour	Modéré	Évolutivité limitée; risque de captation réglementaire ou d'accès inégal

OBJECTIF POLITIQUE	INTERVENTION POLITIQUE	DESCRIPTION	EXEMPLE	INTRUSION SUR LE MARCHÉ	COMPROMIS POTENTIEL
Capacités et contrôles de déploiement	Examen préalable au déploiement	Examen ou accès au modèle avancé avant le déploiement	États-Unis	Modéré	Retards potentiels de mise sur le marché; charge administrative; la participation volontaire peut limiter la couverture
	Réglementation fondée sur les capacités	Réglementations déclenchées par les capacités du modèle	UE	Élevé	Incertitude réglementaire concernant les seuils
	Interdiction de la réplique autonome	Interdiction de certaines formes de réplique autonome ou incontrôlée par les systèmes d'IA	Sans objet	Extrêmement élevée	Pourrait empêcher la mise en place d'applications utiles à l'avenir
	Restrictions relatives à la poursuite autonome d'objectifs	Restrictions applicables aux systèmes d'IA présentant des formes ou un degré spécifiques de poursuite autonome d'objectifs	Sans objet	Extrêmement élevée	Pourrait limiter les applications bénéfiques impliquant des agents
	Restrictions liées à la mise à l'échelle	Restrictions concernant l'augmentation des capacités des modèles, de la puissance de calcul, de la taille ou d'autres intrants de développement	Sans objet	Extrêmement élevée	Peut ralentir les développements scientifiques et commerciaux bénéfiques

OBJECTIF POLITIQUE	INTERVENTION POLITIQUE	DESCRIPTION	EXEMPLE	INTRUSION SUR LE MARCHÉ	COMPROMIS POTENTIEL
Prévention et intervention en cas d'urgence	Exigences de pré-déploiement à haut risque	Respect d'exigences spécifiques, telles que la documentation, la gestion des risques ou les essais, avant le déploiement	UE	Élevé	Charge de mise en conformité importante et obstacles potentiels à l'accès au marché
	Protocole d'intervention en cas d'incident	Mettre en place des procédures, définir les responsabilités et fixer des délais pour réagir aux incidents critiques	États-Unis (Illinois)	Élevé	Risque de surdéclaration ou de classification incohérente des incidents
	Obligation de disposer d'un « kill switch »	Obligation de disposer d'une capacité technique permettant de mettre hors service certains systèmes dotés d'IA	États-Unis (proposition)	Extrêmement élevée	Faisabilité technique; ambiguïté quant aux conditions d'activation
	Pouvoirs d'intervention d'urgence du gouvernement	Pouvoir du gouvernement d'intervenir dans des conditions spécifiques	Sans objet	Extrêmement élevé	Risque d'ingérence excessive du gouvernement; incertitude pour les entreprises
	Interruption ou ralentissement du développement	Suspension temporaire ou ralentissement du développement	Sans objet, mais initiative du secteur	Extrêmement élevée	Risque de conformité inégale à l'échelle internationale; perte d'innovations bénéfiques
	Moratoire sur le développement ou le déploiement	Interdiction de certaines activités de développement ou de déploiement de l'IA	Sans objet	Extrêmement élevée	Coûts élevés en matière d'innovation et de compétitivité

Exemples d'outils politiques disponibles au niveau individuel

OBJECTIF POLITIQUE	INTERVENTION POLITIQUE	DESCRIPTION	EXEMPLE	INTRUSION SUR LE MARCHÉ	COMPROMIS POTENTIEL
Transparence	Transparence de la publicité politique	Divulgarion d'informations sur la publicité politique et indication si le contenu a été généré par l'IA	UE	Faible	Charge administrative pesant sur les campagnes et les plateformes
	Obligations de transparence en matière d'IA	Information des utilisateurs indiquant qu'ils interagissent avec une IA	UE	Faible	Coûts de mise en œuvre et d'interopérabilité
	Étiquetage et tatouage numérique du contenu généré par l'IA	Mention des contenus synthétiques et intégration de signaux de provenance	UE	Faible	Les étiquettes peuvent être supprimées ou contournées
	Signatures numériques	Mécanisme cryptographique ou autre mécanisme de provenance permettant d'authentifier l'origine du contenu	États-Unis (Californie)	Faible	Coûts d'infrastructure, d'interopérabilité et de mise en œuvre
Gestion des risques	Mesures de cybersécurité et exigences en matière de renforcement des modèles	Exigences visant à protéger les systèmes d'IA contre les attaques, la manipulation, les accès non autorisés et autres menaces de sécurité	UE	Élevé	Coûts de mise en conformité élevés ; réduction possible de la flexibilité des modèles
	Détection et signalement des utilisations abusives	Surveillance continue des systèmes déployés afin de détecter et de signaler les utilisations abusives	N/A, mais proposition de Rand	Élevé	Problèmes de confidentialité ; faux positifs ; coûts de surveillance
	Mesures de protection obligatoires contre les utilisations abusives	Mesures de sécurité requises pour prévenir toute utilisation abusive, telle que des activités criminelles ou la fabrication d'armes chimiques et biologiques, avant le déploiement	Sans objet, mais techniques utilisées par l'industrie	Élevé	Fonctionnalité réduite pour certaines utilisations légitimes
	Contrôles d'accès aux matières nécessaires à la fabrication d'armes chimiques ou biologiques	Accès physique restreint aux produits chimiques et aux matériaux utilisés pour la fabrication d'armes chimiques ou biologiques	États-Unis	Extrêmement élevé	Accès restreint pour la recherche légitime

OBJECTIF POLITIQUE	INTERVENTION POLITIQUE	DESCRIPTION	EXEMPLE	INTRUSION SUR LE MARCHÉ	COMPROMIS POTENTIEL
Contrôles de déploiement	Connaissance du client	Vérification de l'identité des utilisateurs accédant à des services ou fonctionnalités spécifiques	États-Unis	Élevé	Préoccupations en matière de confidentialité et d'anonymat
	Modération des contenus	Mesures visant à identifier, restreindre ou supprimer les contenus générés par l'IA qui sont préjudiciables ou interdits	Royaume-Uni	Élevé	Préoccupations liées à la liberté d'expression ; erreurs de modération
	Licences informatiques	Exigences en matière de licence ou d'autorisation régissant l'accès aux ressources de calcul avancées	États-Unis	Extrêmement élevé	Obstacle majeur à l'accès et concentration potentielle des ressources de calcul avancées
	Licences de diffusion des poids	Exigences en matière de licence ou d'autorisation régissant l'accès aux poids de modèles d'IA avancés	États-Unis	Extrêmement élevé	Concentration du marché et accès restreint à la recherche ouverte

Exemples d'outils politiques disponibles au niveau sociétal

OBJECTIF POLITIQUE	INTERVENTION POLITIQUE	DESCRIPTION	EXEMPLE	INTRUSION SUR LE MARCHÉ	COMPROMIS POTENTIEL
Renforcement de la résilience	Programmes d'éducation à l'IA et au numérique	Sensibilisation du public aux systèmes d'IA, à leur utilisation et à leurs impacts, y compris la capacité à identifier et à évaluer de manière critique les contenus générés par l'IA	Finlande	Faible	Impact mesurable limité; investissement public soutenu; horizon temporel long pour les retombées positives
	Initiatives en faveur d'une information fiable et soutien aux médias indépendants	Soutien public en faveur d'écosystèmes d'information fiables et de sources d'information dignes de confiance	Royaume-Uni	Faible	Dépendance vis-à-vis des financements publics; préoccupations potentielles concernant l'influence du gouvernement sur l'information
	Soutien à la vérification des faits	Vérification des informations par des tiers pour lutter contre la désinformation générée par l'IA	Brésil	Faible	Perception d'un parti pris politique; inquiétudes quant à la pérennité
	Modifications de la loi électorale et règles spécifiques aux élections	Règles régissant l'utilisation de contenus générés par l'IA dans un contexte électoral	Brésil Canada	Faible	Risque d'ingérence excessive dans le discours politique et difficultés à définir les contenus interdits
	Programmes de formation continue, de perfectionnement et de reconversion professionnelle	Programmes de formation pour le développement continu des compétences ou pour la reconversion professionnelle	Singapour	Modéré	Coût budgétaire et délai de mise en œuvre long
	Adaptation de l'assurance chômage	Adaptation des systèmes de soutien au revenu face aux suppressions d'emplois liées à l'IA	États-Unis (projet de loi, New Jersey)	Modéré	Coût budgétaire; risque de dissuader un retour rapide à l'emploi
	Programme d'aide à la transition	Soutien ciblé aux travailleurs ayant perdu leur emploi ou menacés de le perdre	Sans objet, mais initiative RAISE	Élevé	Coût budgétaire
	Revenu de base universel	Aide régulière sous forme de revenu de base versée aux travailleurs	Finlande	Extrêmement élevé	Coût budgétaire très élevé; effet sur l'emploi incertain

OBJECTIF POLITIQUE	INTERVENTION POLITIQUE	DESCRIPTION	EXEMPLE	INTRUSION SUR LE MARCHÉ	COMPROMIS POTENTIEL
Protection des droits	Négociation collective	Représentation des travailleurs et mécanisme de négociation collective visant à faire face à l'impact de l'IA sur le lieu de travail	Suède	Modéré	Coût de la main-d'œuvre potentiellement plus élevé
	Protections de l'emploi	Règles imposant aux employeurs de respecter certaines obligations lorsque des changements liés à l'IA ont des répercussions sur l'emploi ou le licenciement	France	Élevé	Coût de la main-d'œuvre potentiellement plus élevé; flexibilité réduite en matière d'ajustement des effectifs
	Législation en matière de protection des consommateurs	Adapter la législation générale sur la protection des consommateurs aux produits et services liés à l'IA	Australie	Élevé	Coût de mise en conformité
	Protection de la vie privée	Règles régissant la collecte, l'utilisation et le traitement des données à caractère personnel	UE	Élevé	Disponibilité réduite des données
	Protection des droits d'auteur	Règles régissant l'utilisation d'œuvres protégées par le droit d'auteur dans le cadre du développement et de l'entraînement de l'IA	UE	Élevé	Disponibilité réduite des données
	Règles anti-discrimination	Règles exigeant des garanties contre la discrimination dans la prise de décision assistée par l'IA	États-Unis (Californie)	Élevé	Augmentation des coûts liés à la mise en conformité et aux litiges
	Droit à une explication	Droit de la personne concernée à recevoir des informations pertinentes sur le rôle de l'IA dans l', dans le cadre de certaines décisions	UE	Élevé	Charge administrative et efficacité potentiellement réduite
	Évaluation de l'impact sur les droits fondamentaux	Évaluation structurelle de la manière dont les systèmes d'IA peuvent affecter les droits fondamentaux et des mesures d'atténuation nécessaires	UE	Élevé	Coût de mise en conformité
	Droit à un contrôle humain	Droit de faire examiner une décision automatisée par une personne dans des circonstances spécifiques	UE	Élevé	Baisse d'efficacité et augmentation des coûts

OBJECTIF POLITIQUE	INTERVENTION POLITIQUE	DESCRIPTION	EXEMPLE	INTRUSION SUR LE MARCHÉ	COMPROMIS POTENTIEL
Recherche et apprentissage	Indicateurs relatifs à l'IA	Indicateurs statistiques destinés à suivre l'adoption de l'IA, ses capacités, ses effets économiques ou ses impacts sociétaux	Corée du Sud	Faible	Coût de maintenance et horizon à plus long terme pour constituer des données longitudinales
	Simulations, prospective et élaboration de scénarios	Exploration structurée des évolutions plausibles de l'IA et de leurs implications sociétales	Royaume-Uni	Faible	Coût d'opportunité
	Observatoires de l'IA et centres de recherche	Institutions publiques chargées de surveiller et d'étudier l'impact de l'IA sur la société et l'économie	Royaume-Uni	Faible	Financement public soutenu ; risque de chevauchement entre les institutions

Exemples d'outils politiques disponibles au niveau international

OBJECTIF POLITIQUE	INTERVENTION POLITIQUE	DESCRIPTION	EXEMPLE	INTRUSION SUR LE MARCHÉ	COMPROMIS POTENTIEL
Coordination	Groupes d'experts scientifiques internationaux	Organismes consultatifs d'experts fournissant des données scientifiques sur l'état actuel des connaissances	ONU	Faible	Formation lente du consensus; difficulté à traduire l'évaluation scientifique en politiques
	Réseau international d'instituts de sécurité	Coopération internationale entre les instituts spécialisés dans la sûreté et la sécurité de l'IA en matière de recherche, d'évaluation et de partage d'informations	Réseau international des instituts de sécurité de l'IA	Faible	Coûts de coordination, divergences entre les mandats nationaux et risques de chevauchements
	Système mondial de signalement des incidents	Mécanisme international de collecte et de partage d'informations sur les incidents liés à l'IA	OCDE	Faible	Préoccupations liées à la confidentialité et à la qualité des données ; la participation pourrait être inégale
	Rapports internationaux sur la transparence	Approches communes en matière de communication d'informations sur les systèmes d'IA, les risques et les pratiques de sécurité	G7 HAIP	Modéré	La mise en place d'un accord sur des exigences communes en matière de reporting peut prendre beaucoup de temps ; charge administrative liée au reporting
	Normes internationales	Accord sur une terminologie commune, des spécifications techniques, des pratiques d'évaluation et des règles de procédure	UIT	Modéré	L'élaboration des normes prend du temps; celles-ci risquent d'être obsolètes au moment de leur adoption
	Traités internationaux	Accord sur des règles communes concernant la conception, le développement, le déploiement et l'utilisation	Conseil de l'Europe	Modéré à élevé	Difficiles à négocier, à ratifier et à modifier; leur mise en œuvre peut varier d'une juridiction à l'autre

Annexe 4

Étapes pour l'élaboration de politiques publiques face aux risques liés à l'IA

1. Identifier et définir le risque sur la base de données probantes

- Caractériser:
 - Préjudices potentiels et parties prenantes concernées
 - Probabilité, gravité et réversibilité des préjudices potentiels
 - Type et degré d'incertitude
 - Horizon temporel (court, moyen ou long terme)
 - Niveau du risque (modèle, individu, société, international)
 - Données existantes et lacunes dans les connaissances

2. Évaluer si une intervention des pouvoirs publics est justifiée

- Prendre en compte:
 - Les défaillances du marché
 - Asymétries d'information
 - Externalités
 - Cadres juridiques et réglementaires existants
 - Initiatives existantes du secteur
- Si une intervention n'est pas justifiée à l'heure actuelle, poursuivre la surveillance et réévaluer la situation à mesure que de nouvelles données apparaissent.

3. Définir les objectifs politiques

- Préciser les résultats escomptés de toute intervention
- S'assurer que les objectifs sont clairs, mesurables et réalisables

4. Identifier les instruments politiques potentiels

- Explorer l'ensemble des outils disponibles, notamment:
 - Les instruments juridiques contraignants

- Les instruments de gouvernance non contraignants
- Les instruments expérimentaux
- Les instruments programmatiques
- Les instruments d'incitation économique
- Instruments institutionnels

5. Évaluer les options politiques

- Évaluez chaque option à l'aune de critères pertinents, tels que:
 - Efficacité
 - Coûts et avantages
 - Faisabilité et applicabilité de l'
 - Flexibilité et adaptabilité de l'
 - Acceptabilité sociale et politique
 - Conséquences imprévues potentielles
 - Compromis

6. Choisir l'approche politique privilégiée

- Choisir l'option, ou la combinaison d'options, qui permet le mieux d'atteindre les objectifs politiques tout en minimisant les coûts et les effets indésirables (cela peut inclure le maintien du statu quo)

7. Élaborer un plan de mise en œuvre

- Définir les rôles, les responsabilités, les délais et les ressources
- Anticiper les difficultés de mise en œuvre et les stratégies d'atténuation
- Identifier les indicateurs, les seuils de déclenchement et les mesures d'urgence

8. Assurer le suivi, évaluer et adapter

- Suivre les résultats et la conformité
- Évaluer si les objectifs sont atteints
- Réviser l'intervention à mesure que les données, les technologies et les risques évoluent

