Un nouveau cadre du NIST fournit des orientations sur la gouvernance et la gestion du risque en matière d'IA

14 mars 2023

Le 26 janvier 2023, le National Institute of Standards and Technology (NIST) des États-Unis a publié le document intitulé *Artificial Intelligence Risk Management Framework* (cadre de gestion du risque en matière d'intelligence artificielle, le « cadre du NIST »). Reconnaissant la complexité sans égale des outils et des systèmes d'IA et les vastes enjeux qu'ils posent, ce cadre contient des orientations pour la gestion des risques liés à leur conception, à leur développement, à leur déploiement et à leur utilisation, comme la prise de décisions biaisées, et propose de bonnes pratiques en la matière.

Le cadre du NIST fournit des orientations utiles pour la gouvernance et la gestion du risque en matière d'IA à l'heure où les organisations canadiennes se préparent à l'entrée en vigueur de lois encadrant les outils de prise de décision automatisée (en septembre 2023 au Québec). Sa publication coïncide aussi avec la reprise des débats de la Chambre des communes sur le projet de loi C-27 et les lois qu'il propose, la *Loi* sur la protection de la vie privée des consommateurs (LPVPC) et la Loi sur l'intelligence artificielle et les données (LIAD).

Même en l'absence de lois et de règlements spécifiques encadrant l'IA, le cadre du NIST peut aider les organisations à mettre en place une structure de gouvernance et des contrôles destinés à favoriser une conception, un développement, un déploiement et une utilisation responsables des systèmes d'IA et à atténuer les risques liés à la conformité connexes (lesquels peuvent découler de lois existantes, comme celles encadrant la protection de la vie privée, l'emploi et la protection des droits de la personne).

En quoi consiste le cadre du NIST?

Le NIST relève du Department of Commerce des États-Unis et a pour mission de promouvoir l'innovation et la compétitivité industrielle américaines en faisant progresser la métrologie et les normes et technologies connexes dans le but de renforcer la sécurité économique et d'améliorer la qualité de vie.

L'*Artificial Intelligence Risk Management Framework* est un cadre d'application volontaire qui outille les organisations afin qu'elles puissent évaluer, classifier, communiquer et gérer les risques que présentent les systèmes d'IA. Ce document présente aussi sept qualités d'un système d'IA digne de confiance.

En marge de la publication de son cadre, le NIST a aussi publié plusieurs documents d'accompagnement et une vidéo explicative :

- un guide présentant des processus que les organisations peuvent adapter pour obtenir les quatre résultats que produit le cadre : gouverner, cartographier, mesurer et gérer les risques associés aux systèmes d'IA;
- un plan directeur, qui dresse une liste des activités que le NIST pourrait entreprendre, seul ou avec des parties prenantes des secteurs privé et public, pour faire évoluer son cadre. Par exemple, comme les organisations doivent définir leur tolérance au risque pour utiliser ce dernier, le NIST souhaite leur proposer des méthodes à suivre pour relever leur tolérance au risque à un niveau raisonnable.
- des tableaux de concordance décrivant les liens entre le cadre et des normes ou orientations antérieures sur l'IA. Deux projets de tableaux de concordance ont déjà été publiés. Ils concernent i) la norme ISO/IEC 23894:2023 sur l'IA et ii) une illustration des liens entre les qualités d'un système d'IA digne de confiance selon le cadre du NIST et celles que l'on retrouve dans la recommandation de l'OCDE sur l'IA, le projet de règlement de l'UE sur l'IA, le U.S. Executive Order 13960 (*Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government*), et le Blueprint for an AI Bill of Rights de la Maison-Blanche.

Comment le cadre du NIST définit-il les systèmes d'IA?

Le cadre propose une définition de « système d'IA » adaptée de la recommandation de l'OCDE sur l'IA et de la norme ISO/IEC 23894:2023 sur l'IA, qui a des points en commun avec la définition de la LIAD :

Cadre du NIST	Article 2 de la LIAD
[Traduction] *Système technique ou automatisé qui peut, pour un ensemble donné d'objectifs, générer des produits tels que des prédictions, des recommandations ou des décisions ayant une incidence sur des environnements réels ou virtuels. Les systèmes d'IA sont conçus pour fonctionner avec des degrés variables d'autonomie.*	*Système technologique qui, de manière autonome ou partiellement autonome, traite des données liées à l'activité humaine par l'utilisation d'algorithmes génétiques, de réseaux neuronaux, d'apprentissage automatique ou d'autres techniques pour générer du contenu, faire des prédictions ou des recommandations ou prendre des décisions.*

Dans le cadre du NIST comme dans la LIAD, l'intention semble être de couvrir un large éventail de systèmes. Si les deux définitions renvoient aux mêmes types de produits, à savoir des décisions, des recommandations et des prédictions, la LIAD mentionne aussi la production de *contenu*. Contrairement à la LIAD, cependant, le cadre du NIST ne mentionne pas de types précis de technologies, dans le souci sans doute de n'avoir

aucun parti pris technologique. Par ailleurs, les deux définitions reconnaissent que les systèmes d'IA peuvent comporter des degrés d'autonomie variables. Cela concorde avec la définition de système décisionnel automatisé de la LPVPC, mais contraste avec le projet de loi 64 du Québec, qui réglemente l'utilisation de renseignements personnels afin que soit rendue une décision fondée *exclusivement* sur un traitement automatisé (paragraphe 12.1.).

Encadrer le risque : évaluer, classifier, communiquer et gérer les risques que posent les systèmes d'IA

Le cadre du NIST présente des orientations sur la gestion du risque visant à atténuer les incidences négatives des systèmes d'IA. Contrairement aux projets de loi sur l'IA à l'étude au Canada et dans l'UE, le cadre du NIST ne propose pas de classification des systèmes d'IA fondée sur leur incidence potentielle ou le risque qui leur est associé, et il ne fixe pas de critères précis pour l'évaluation du risque. Par contraste, la LIAD exige que les organisations visées évaluent si leur système d'IA est un « système à incidence élevée » selon les critères de la réglementation (article 7). En présence d'un système à incidence élevée, il y aurait des obligations d'atténuation du risque, de transparence et d'explicabilité, de même qu'une obligation d'aviser le ministre si l'utilisation du système entraînait, ou entraînerait vraisemblablement, un préjudice important (articles 8, 9, 11 et 12).

Le cadre du NIST fournit des orientations aux organisations pour les aider à comprendre et à gérer les risques, les incidences et les préjudices associés à l'IA. Il donne les exemples de préjudices suivants :

- les préjudices causés aux personnes, qui comprennent les préjudices causés à l'individu (atteinte aux droits civils), les préjudices causés à un groupe (comme la discrimination) et les préjudices sociétaux (comme les barrières à l'éducation);
- les préjudices causés aux activités ou à la réputation d'une organisation et ceux qui découlent d'atteintes à la sécurité ou de pertes financières;
- les préjudices causés à un écosystème, comme le système financier mondial ou les ressources naturelles.

Fait intéressant, ces exemples ratissent plus large que la définition de préjudice de la LIAD, qui se concentre sur le préjudice causé aux personnes (la définition du paragraphe 5(1) se lit comme suit : préjudice physique ou psychologique subi par un individu, dommage à ses biens ou perte économique subie par un individu).

Le cadre du NIST note que, pour encadrer les risques, les organisations devront relever certains défis, en particulier les suivants :

- Évaluation du risque : Le cadre commence par proposer que les risques posés par les systèmes d'IA soient mesurés au moyen d'une matrice classique, selon laquelle le risque dépend de deux choses : i) l'incidence négative (ou l'ampleur du préjudice) qui découlerait de la concrétisation des circonstances ou de l'événement et ii) la probabilité de concrétisation. Les risques ne sont pas nécessairement tous prévisibles au moment de la conception ou de la mise en

place d'un système d'IA et, s'ils le sont, ils peuvent être difficiles à mesurer. Le type d'évaluation qu'ils exigent peut aussi varier selon le stade du cycle de vie de l'IA auquel ils se présentent. Les organisations auront intérêt à adopter une approche souple de la gestion du risque pour pouvoir réagir adéquatement si de nouveaux risques devaient se poser.

- Tolérance au risque :Le cadre n'énonce pas de règles en matière de tolérance au risque associé aux systèmes d'IA. Il présente cependant quelques points sur lesquels peuvent s'appuyer les organisations pour déterminer leur degré de tolérance. La tolérance au risque d'une organisation, en ce qui concerne son utilisation d'un système d'IA, variera selon la finalité du système et selon les politiques et les normes établies par diverses parties (p. ex., le propriétaire du système, ses utilisateurs et des décideurs gouvernementaux et non gouvernementaux). De plus, la tolérance au risque sera vraisemblablement appelée à évoluer au fil du temps et au cours du cycle de vie du système d'IA.
- Détermination de l'ordre de priorité des risques : Reconnaissant qu'il est utopique d'espérer éliminer tous les risques, le cadre propose des principes à suivre pour déterminer lesquels sont les plus menaçants. Le développement d'une forte culture de gestion du risque et l'adoption de protocoles de triage efficaces permettront d'affecter des ressources à la gestion des risques les plus graves en premier lieu. Une organisation devrait par ailleurs se demander quels facteurs propres à sa situation devraient peser davantage que d'autres dans la détermination de l'ordre de priorité. Il conviendrait, par exemple, de classer comme prioritaire un système d'IA qui traite des renseignements personnels ou qui gère de grands ensembles de données. Enfin, une fois qu'une organisation a déterminé l'ordre de priorité des risques associés à un système d'IA et qu'elle a agi pour les réduire le plus possible, elle doit évaluer le risque résiduel pour les utilisateurs finaux et être convaincue qu'il est tolérable.
- Intégration à la gestion globale du risque : Les organisations doivent éviter de traiter isolément les risques liés aux systèmes d'IA. Elles doivent plutôt les intégrer à leur stratégie globale de gestion du risque. En effet, certains risques associés aux systèmes d'IA, comme les questions de confidentialité et de cybersécurité, se rencontrent aussi dans des processus comme l'élaboration de logiciels ou la gestion des données. D'autres sont néanmoins propres aux systèmes d'IA; un travail considérable peut devoir être accompli pour situer ceux-là dans les pratiques de gestion du risque de l'organisation.

Les qualités d 'un système d 'IA digne de confiance

Le cadre du NIST présente sept qualités d'un système d'IA digne de confiance. Toutes ne s'appliquent pas dans la même mesure à tous les systèmes d'IA. De plus, comme les systèmes d'IA ne vivent pas en vase clos, l'importance de chaque qualité peut varier selon l'utilisation qui est faite d'un système et son contexte d'utilisation. L'importance de chaque qualité dépend aussi des données sur lesquelles repose le système, de ce qu'il produit, des décisions de son créateur et de la mesure dans laquelle des humains interagissent avec lui ou le supervisent.

Voici, selon le cadre du NIST, les qualités d'un système d'IA digne de confiance :

1) Redevabilité et transparence : Les utilisateurs d'un système d'IA devraient avoir accès à des informations d'un niveau approprié à son sujet. Ce niveau dépend de

la manière dont le système est utilisé et de son contexte d'utilisation, du risque qui y est associé et du stade qu'il a atteint dans son cycle de vie, entre autres choses.

Ces qualités correspondent globalement aux exigences de transparence que fixe la LIAD pour les systèmes à incidence élevée. Plus précisément, la LIAD exige que les personnes qui rendent disponible un système à incidence élevée ou qui en gèrent l'exploitation publient, sur un site Web accessible au public, une description en langage clair du système. Cette description doit comprendre les éléments énumérés dans la LIAD, selon des modalités fixées par règlement (article 11).

Les projets de loi 64 et C-27 fixent aussi des exigences de transparence en ce qui concerne la prise de décision automatisée. Par exemple, selon le projet de loi 64, toute personne qui exploite une entreprise et qui utilise des renseignements personnels afin que soit rendue une décision fondée exclusivement sur un traitement automatisé de ceux-ci doit en informer la personne concernée (article 12.1.) La personne concernée a également le droit de présenter des observations à un membre du personnel de l'organisation en mesure de réviser la décision. L'organisation doit aussi, à la demande de la personne concernée, l'informer : 1) des renseignements personnels utilisés pour rendre la décision; 2) des raisons, ainsi que des principaux facteurs et paramètres, ayant mené à la décision; et 3) de son droit de faire rectifier les renseignements personnels utilisés pour rendre la décision.

2) Validité et fiabilité : Un système d'IA doit remplir sa tâche correctement et avec constance au fil du temps et dans les diverses circonstances pour lesquelles il a été conçu. Cette qualité sous-tend aussi les cinq autres énumérées ci-après.

La LIAD n'impose pas explicitement d'obligations à cet égard.

3) Sûreté :Les organisations doivent prioriser la sûreté à chaque stade du cycle de vie de l'IA et ne doivent pas concevoir ni utiliser de systèmes d'IA qui, « [traduction] dans des conditions définies, conduisent à un état où la vie humaine, la santé humaine, la propriété ou l'environnement est mis en danger ». En plus de concevoir, de développer et de déployer des systèmes d'IA de manière responsable, les organisations doivent fournir des informations claires sur la manière d'interagir avec eux en toute sécurité, et s'assurer qu'elles respectent les normes de sécurité en vigueur.

Cette qualité se compare aux principes de la LIAD concernant l'atténuation du risque de préjudice pour les systèmes à incidence élevée.

4) Sécurité et résilience : Dans le monde infiniment connecté d'aujourd'hui, la cybersécurité doit être au cœur des préoccupations aux étapes de la conception et du déploiement des systèmes d'IA. Ces systèmes doivent pouvoir résister à des événements indésirables et inattendus et, quand ce n'est pas possible, être conçus pour faire défaut en toute sécurité.

La LIAD ne traite pas directement de ces qualités, mais les systèmes d'IA qui n'intègrent pas la sécurité et la résilience posent un risque de préjudice et de résultats biaisés, lesquels sont des aspects fondamentaux du projet de loi.

5) Explicabilité et interprétabilité : Toute personne qui interagit avec un système d'IA doit pouvoir en comprendre la raison d'être et l'incidence.

Ces principes vont de pair avec les principes de la redevabilité et de la transparence et concordent avec les exigences d'explicabilité de la LIAD, du projet de loi 64 et de la LPVPC.

6) Confidentialité améliorée : La confidentialité doit être au cœur de l'élaboration des systèmes d'IA. Les technologies d'amélioration de la confidentialité, la dépersonnalisation et l'agrégation de données sont tous des outils utiles pour développer des systèmes d'IA à confidentialité améliorée. Cela dit, lorsque les données sont peu nombreuses ou qu'elles sont incomplètes pour d'autres raisons, ces stratégies peuvent entraîner une perte d'exactitude des systèmes d'IA.

Au Canada, les exigences relatives à la confidentialité seraient fixées par les lois sur la protection des renseignements personnels plutôt que par la LIAD.

7) Équité et gestion des biais préjudiciables : Les systèmes d'IA étant le produit des humains qui les ont créés et des données qu'ils ont utilisées pour ce faire, des mesures doivent être prises pour réduire et gérer les biais aux stades de la conception et de l'utilisation. Ces qualités se comparent aux principes qu'énonce la LIAD en ce qui concerne le risque de résultats biaisés pour les systèmes à incidence élevée.

Points à retenir

Même en l'absence de lois sur l'intelligence artificielle au Canada, toute organisation canadienne qui conçoit, développe, déploie ou utilise des systèmes d'IA devrait se doter d'un cadre pour atténuer les diverses catégories de risques (observation de la loi, éthique, exploitation, réputation, etc.) que peuvent poser ces activités. Cela peut concerner tout un éventail d'organisations qui ne sont pas nécessairement des entreprises de technologie, puisque désormais, les outils de TI comportent la plupart du temps un volet d'IA.

Le cadre du NIST peut aider ces organisations à établir un cadre de gouvernance en matière d'IA. Celles qui auront déjà pris des mesures d'atténuation des risques auront une longueur d'avance quand seront adoptées des lois encadrant l'IA et des lois sur la protection des renseignements personnels régissant des systèmes d'IA en particulier.

Si vous avez des questions sur le nouveau cadre du NIST et sur la manière dont il peut aider votre organisation au sujet de la gouvernance et de la gestion du risque en matière d'IA, n'hésitez pas à communiquer avec l'une ou l'autre des personnes ci-dessous.

Par

Marc Vani

Services

Cybersécurité, respect de la vie privée et protection des renseignements personnels, Technologies, Intelligence artificielle (IA)

BLG | Vos avocats au Canada

Borden Ladner Gervais S.E.N.C.R.L., S.R.L. (BLG) est le plus grand cabinet d'avocats canadien véritablement multiservices. À ce titre, il offre des conseils juridiques pratiques à des clients d'ici et d'ailleurs dans plus de domaines et de secteurs que tout autre cabinet canadien. Comptant plus de 725 avocats, agents de propriété intellectuelle et autres professionnels, BLG répond aux besoins juridiques d'entreprises et d'institutions au pays comme à l'étranger pour ce qui touche les fusions et acquisitions, les marchés financiers, les différends et le financement ou encore l'enregistrement de brevets et de marques de commerce.

blg.com

Bureaux BLG

Calgary

Centennial Place, East Tower
520 3rd Avenue S.W.
Calgary, AB, Canada
T2P 0R3

T 403.232.9500
F 403.266.1395

Ottawa

World Exchange Plaza
100 Queen Street
Ottawa, ON, Canada
K1P 1J9

T 613.237.5160
F 613.230.8842

Vancouver

1200 Waterfront Centre
200 Burrard Street
Vancouver, BC, Canada
V7X 1T2

T 604.687.5744
F 604.687.1415

Montréal

1000, rue De La Gauchetière Ouest
Suite 900
Montréal, QC, Canada
H3B 5H4

T 514.954.2555
F 514.879.9015

Toronto

Bay Adelaide Centre, East Tower
22 Adelaide Street West
Toronto, ON, Canada
M5H 4E3

T 416.367.6000
F 416.367.6749