

Construire une assurance de l'IA interopérable entre le Royaume-Uni et le Canada

Préparé par

Connor Attridge, Chercheur associé, Centre for Emerging
Technology and Security – The Alan Turing Institute

Henri Vilandré, Conseiller principal en politiques, Mila

Cette analyse a été élaborée en collaboration avec le Centre for Emerging Technology and Security ([CETaS](#)) de l'Institut Alan Turing. Une copie est également disponible sur [leur site web](#).

Avertissement : Les opinions et points de vue exprimés dans ce document n'engagent que leurs auteurs et ne reflètent pas la politique ou la position officielle des instituts éditeurs.

5 octobre 2026

Construire une assurance de l'IA interopérable entre le Royaume-Uni et le Canada

Alors que les systèmes d'IA sont de plus en plus utilisés au sein des gouvernements, des infrastructures essentielles et des environnements de sécurité, l'assurance de l'IA peut aider les organisations à se conformer aux exigences réglementaires et à déployer l'IA de manière sécuritaire. Des [incidents récents](#) ont montré que les agents d'IA avancés peuvent agir au-delà du cadre prévu – même lors de séances d'évaluation contrôlées. Cela souligne la nécessité d'évaluer et de valider non seulement la fiabilité des modèles sous-jacents, mais aussi les conditions d'exploitation et les processus institutionnels liés à l'utilisation d'un système.

La visite du premier ministre canadien Mark Carney en Europe, en septembre 2026, a mis en évidence son [ambition](#) de renforcer la coopération avec des partenaires partageant les mêmes valeurs au sujet de l'IA. Lors de son discours devant le Parlement européen, il a proposé de mener des travaux conjoints sur la sécurité, les normes et la transparence de l'IA. Au cours de cette même visite, le Royaume-Uni et le Canada ont accepté de [renforcer leur coopération](#) sur l'IA sécuritaire et responsable. Le Canada et le Royaume-Uni – des alliés dans le cadre du Groupe des Cinq qui entretiennent des relations bien établies en matière d'échange d'informations – disposent d'une base solide pour le faire. Travailler pour une assurance de l'IA interopérable – où les éléments probants liés à l'assurance peuvent être partagés, compris et utilisés dans différentes juridictions – offre une façon de concrétiser cette ambition politique.

Le fondement politique de la coopération existe déjà. En juin 2026, le Royaume-Uni et le Canada se sont [engagés](#) à développer une politique et des normes en matière d'IA et à renforcer la coopération dans le domaine des systèmes de pointe au service de la sécurité nationale. Cela repose sur la coopération en matière de [capacités de calcul en IA](#), de [recherche sur la sécurité](#) et d'[échanges entre chercheur·euse·s](#). La question pratique consiste à traduire ces engagements en décisions quant à savoir si, et sous quelles conditions, les preuves provenant de partenaires peuvent satisfaire les exigences nationales d'assurance.

Pour les puissances moyennes, la souveraineté en IA dépend en partie du choix des technologies et des fournisseurs d'IA. La coopération sur l'assurance peut aider les deux pays à évaluer ces options. Nous proposons une approche pratique de la coopération entre le Royaume-Uni et le Canada sur l'assurance de l'IA avec trois objectifs principaux :

1. **Faciliter un déploiement sécuritaire**, en mettant en place un processus concerté pour l'adoption de l'IA au sein de secteurs critiques et à enjeux élevés.

2. **Préserver et renforcer l'autonomie des puissances moyennes**, où les deux nations établissent ensemble les critères déterminants pour une IA sécuritaire tout en conservant le contrôle sur les décisions de déploiement sur le plan national.
3. **Construire une confiance institutionnelle grâce à un travail d'évaluation conjoint**, qui détermine les moments où la coopération peut s'étendre à des domaines plus sensibles.

En s'inspirant de l'expérience de l'aviation civile, le Royaume-Uni et le Canada devraient évaluer les cas où les conclusions d'une évaluation réalisée par un partenaire peuvent satisfaire aux exigences nationales d'assurance.

Assurance et institutions

L'assurance de l'IA, au sens large, désigne la manière dont une organisation s'assure qu'un système d'IA fait ce qu'il est censé faire, de manière [sécuritaire et légale](#), dans le contexte prévu. Au-delà de cette définition large, le vocabulaire est contesté. La [recherche CETaS](#) sur le marché britannique de la défense et de la sécurité a révélé que l'assurance « signifie différentes choses à différentes personnes », allant du test technique étroit à des processus de conformité en passant par des définitions plus holistiques. S'entendre sur des définitions de travail de l'assurance, de la vérification et de la validation aiderait les évaluateur-riche-s du Royaume-Uni et du Canada à clarifier ce que les tests effectués par les partenaires permettent de déterminer et quelles preuves supplémentaires sont nécessaires pour justifier un déploiement national.

Deux éléments liés à l'assurance sont pertinents. En premier lieu, l'assurance s'applique à un système dans un contexte de déploiement donné. Les éléments de preuve requis dépendent du système, de son objectif, des conditions d'exploitation et des conséquences d'une défaillance. En deuxième lieu, les méthodes permettant d'assurer la sécurité de ces systèmes évoluent elles-mêmes. Les tests d'évaluation comparative (benchmarking) s'infiltrant dans les données d'entraînement, les bacs à sable sont soit trop restrictifs pour être réalistes, soit trop permissifs pour être sûrs, et les diverses approches d'évaluation de la vulnérabilité (red teaming) révèlent des types de défaillances distincts.

Les méthodes de mesure sur lesquelles repose l'assurance sont cependant de plus en plus structurées. En janvier 2026, le US National Institute of Standards and Technology (NIST, Institut national américain des normes et des technologies) a publié un [projet de lignes directrices](#) pour les évaluations comparatives automatisées des modèles de langage et des

agents d'IA. L'International Network for Advanced AI Measurement, Evaluation and Science (NAAIMES, Réseau international pour la mesure, l'évaluation et la science en IA avancée), dont le Canada et le Royaume-Uni sont membres, a depuis développé des [pratiques consensuelles](#) pour les évaluations automatisées et élabore des [approches communes](#) pour les tests effectués par des tiers. Le Canada et le Royaume-Uni ont déjà participé à des [travaux conjoints en réseau](#) sur les évaluations agentiques, tandis que le [cadre Inspect](#) et les [recommandations techniques](#) en données ouvertes de l'UK AI Security Institute (AIS, Institut britannique pour la sécurité de l'IA) permettent de réutiliser les outils d'évaluation dans différentes organisations.

Cela montre qu'il existe une base de coopération, mais ne justifie pas le remplacement des différent-e-s évaluateur-ice-s nationaux-ales par une seule entité partagée. Le fait de s'appuyer sur un-e seul-e évaluateur-ice ou méthodologie risque de reproduire les mêmes angles morts d'un système à l'autre.

Pourquoi le Royaume-Uni et le Canada ?

Le Royaume-Uni et le Canada sont particulièrement bien placés pour tester ceci puisqu'ils partagent la même position stratégique à titre de puissance moyenne et des capacités complémentaires. Parmi les puissances moyennes dans le domaine de l'IA, le Royaume-Uni a développé et établi une capacité d'évaluation de l'IA par le biais d'AISI. Le système d'évaluation plus récent du Canada couvre plusieurs institutions et bénéficie de nouveaux investissements publics. Une nouvelle stratégie canadienne de l'IA, [L'IA pour tous](#), par exemple, prévoit un montant additionnel de 50 M\$ CA pour augmenter la capacité de l'Institut canadien de la sécurité de l'intelligence artificielle à identifier les risques émergents et à mener des évaluations transparentes.

La logique de cette coopération ne repose pas seulement sur leur statut d'alliés ou sur l'idée d'une équivalence directe entre leurs institutions. Les deux pays partagent un [intérêt stratégique](#) à conserver un contrôle national sur la manière dont les systèmes d'IA, développés à l'intérieur d'un écosystème technologique dominé en grande partie par les États-Unis et la Chine, sont évalués et déployés. Leur partenariat permet de vérifier concrètement si les puissances moyennes peuvent s'appuyer sur leurs forces respectives sans pour autant renoncer à leur souveraineté nationale en matière d'approbation, d'approvisionnement ou de déploiement.

Par où pouvons-nous commencer ?

Tous les cas d'usage de l'IA ne requièrent pas un mécanisme bilatéral. Pour certains usages commerciaux à faible risque, un processus formel pour partager et réviser des éléments de preuve entraînerait des coûts supérieurs aux avantages procurés. La défense est une priorité énoncée pour la coopération Royaume-Uni-Canada : leur [déclaration commune](#) de juin 2025 cerne les contours d'une collaboration en IA de pointe en matière de sécurité nationale et les deux pays entretiennent des relations étroites en défense et en renseignement dans le cadre d'alliances comme le Groupe des Cinq. Certaines applications en défense présentent cependant des contraintes de classification et d'exploitation susceptibles de compliquer la mise en place d'un premier projet pilote.

Une infrastructure nationale essentielle offre plusieurs points de départ prometteurs. Elle comporte des applications ayant des implications directes en matière de défense et de sécurité et des exigences opérationnelles et réglementaires établies, permettant d'évaluer les éléments de preuve. Les deux pays renforcent également leur contrôle de la cybersécurité : la [Loi sur la protection des cybersystèmes essentiels](#) du Canada a reçu la sanction royale en juin 2026, avec une mise en œuvre progressive, tandis que le [Cyber Security and Resilience Bill](#) (projet de loi sur la cybersécurité et la résilience) du Royaume-Uni est actuellement examiné par la Chambre des Lords. Un projet pilote pourrait s'appuyer sur ce contrôle sectoriel, en collaboration avec les opérateur-riche-s et les autorités de régulation concernées, afin d'évaluer dans quelle mesure les éléments de preuve fournis par les partenaires peuvent soutenir les décisions liées au déploiement de l'IA.

Les risques varient selon les applications destinées aux infrastructures nationales essentielles et les secteurs. Par exemple, un modèle de prévision du réseau électrique qui conseille les opérateur-riche-s présente un problème d'assurance différent d'un-e contrôleur-euse en traitement de l'eau qui agit seul-e. Les deux gouvernements pourraient sélectionner une application délimitée dont les conditions d'exploitation, les modes de défaillance et la supervision humaine sont suffisamment maîtrisés pour tester la confiance transfrontalière. Ce projet pilote devrait se concentrer sur un cas d'usage où les exigences en matière d'assurance de l'IA justifient un travail conjoint et où les preuves nécessaires peuvent être partagées et évaluées. Cela peut être réalisé sans avoir à commencer par les cas d'usage les plus sensibles ou à haut risque.

Leçons de l'aviation

L'assurance transfrontalière n'est pas nouvelle ou propre à l'IA. L'aviation fait partie des secteurs qui font face à des problèmes comparables, où l'on doit concilier autorité nationale et exigences de certification. Les ententes de sécurité bilatérales en aviation permettent à une autorité de s'appuyer sur des conclusions techniques précises, produites par un partenaire de confiance et à l'intérieur d'un périmètre convenu, tout en conservant sa propre responsabilité réglementaire.

Le Royaume-Uni et le Canada fonctionnent déjà de cette manière. Un [arrangement de travail](#) entre l'autorité d'aviation civile britannique et Transport Canada existe depuis le 1^{er} janvier 2021 et prévoit la reconnaissance réciproque des résultats concernant la navigabilité et la maintenance. En juin 2026, les deux autorités de régulation [l'ont étendu](#) à une reconnaissance mutuelle des évaluations périodiques réalisées sur simulateur de vol.

Par ailleurs, la matrice des risques pour la sécurité de l'US Federal Aviation Administration (FAA, Administration fédérale de l'aviation aux États-Unis) adapte l'intensité des vérifications au niveau de risque, en classant les défaillances potentielles selon leur gravité et leur probabilité : cas les plus faibles (acceptables), modérés (acceptables avec mesures d'atténuation) et élevés (inacceptables).

L'assurance en IA doit examiner à la fois le modèle et les conditions dans lesquelles il est utilisé. Les environnements incertains ou conflictuels rendent cette évaluation plus difficile. Une [recherche récente de CETaS](#) sur l'assurance comportementale de l'IA agentique met l'accent sur les actions et les interactions des agents, et l'efficacité du contrôle. Des rapports d'évaluation partagés devraient donc décrire les méthodes de test utilisées, les outils disponibles au système, les permissions ainsi que les contrôles organisationnels en place. Cela aiderait les évaluateur-ice-s nationaux-ales à déterminer si les conclusions s'appliquent à l'usage envisagé.

Tableau 1. Matrice des risques pour la sécurité selon l'Administration fédérale de l'aviation (FAA)

Probabilité de risque	Gravité du risque				
	Catastrophique (A)	Dangereux (B)	Majeur (C)	Mineur (D)	Négligeable (E)
Fréquent (5)	5A	5B	5C	5D	5E
Occasionnel (4)	4A	4B	4C	4D	4E
Éloigné (3)	3A	3B	3C	3D	3E
Improbable (2)	2A	2B	2C	2D	2E
Extrêmement improbable (1)	1A	1B	1C	1D	1E

Faible (acceptable)	Modéré (acceptable avec mesures d'atténuation)	Élevé (inacceptable)
---------------------	--	----------------------

Source : Adapté de la FAA, « Safety Management Systems for Aviation Service Providers » (Systèmes de gestion de la sécurité pour les fournisseurs de services d'aviation), Circulaire d'information, 21 mai 2024.

Un modèle de confiance inspiré des feux de circulation

L'approche en matière de sécurité aérienne peut aussi être appliquée à l'assurance de l'IA, en s'appuyant sur trois niveaux de confiance dans l'assurance fournie par le partenaire. La question est de savoir à quelles exigences nationales d'assurance les preuves fournies par le partenaire peuvent répondre. En premier lieu, nous proposons d'appliquer cette approche aux applications délimitées relatives aux infrastructures essentielles :

1. **Vert (plus grande confiance).** Des systèmes délimités dans des environnements contrôlés, comme un système de maintenance prédictive avec un contrôle humain et des vérifications de sécurité existantes. Les évaluations provenant du partenaire peuvent être acceptées, avec un nombre de répétitions limitées et un rapport établi d'un commun accord, lorsqu'elles couvrent l'usage national prévu.
2. **Jaune (confiance conditionnelle).** Les systèmes opérant dans des limites définies, mais exposés à des conditions variables, comme les algorithmes de négociation assortis de limites d'exposition ou des levés maritimes supervisés. Les preuves fournies par le partenaire servent de base pour l'approbation nationale, complétées par de nouveaux essais ciblés. Un accès sécurisé aux poids des modèles, aux données ou à d'autres éléments techniques pertinents peut être nécessaire.
3. **Rouge (évaluation nationale requise).** Les applications pour lesquelles une défaillance pourrait entraîner des conséquences graves ou des responsabilités nationales qui requièrent une évaluation indépendante comme la remise en état, soutenue par l'IA, du réseau électrique national pendant une cyberattaque. Les preuves fournies par le partenaire demeurent informatives, mais une évaluation et une approbation nationale indépendantes sont requises.

Tableau 2. Illustration des niveaux de confiance envers l'assurance fournie par le partenaire

Environnement d'exploitation	Complexité du système		
	Intégré/adaptatif	Connecté	Limité
Contesté/conflictuel	Remise en état, soutenue par l'IA, du réseau électrique national pendant une cyberattaque		
Variable		Levés maritimes autonomes supervisés	
Contrôlé/prévisible		Algorithmes de négociation avec limites d'exposition prédéfinies	Système de maintenance prédictive assisté par l'IA

Plus grande confiance	Confiance conditionnelle	Évaluation nationale requise
-----------------------	--------------------------	------------------------------

Note : Les catégorisations sont à titre illustratif et dépendent du déploiement, des contrôles et des conséquences de la défaillance. Contrairement aux catégories de risque en aviation, le rouge dénote des bases insuffisantes pour s'appuyer uniquement sur l'assurance fournie par le partenaire. L'autorité décisionnelle nationale est préservée tout au long du processus.

Source : Analyse des auteur·rice·s

Un projet pilote devrait tester ces distinctions dans la pratique : quelles conclusions peuvent être acceptées, lesquelles requièrent une validation nationale et quels jugements doivent demeurer nationaux ? Des expériences dans des applications à faible risque peuvent bâtir la confiance institutionnelle et créer des conditions pour une coopération plus large.

Prochaines étapes

Deux étapes peuvent être prises pour tester la confiance transfrontalière :

1. Identifier un projet commun d'infrastructure essentielle.

Les deux gouvernements devraient sélectionner un cas d'usage ciblé et peu sensible d'une infrastructure essentielle et l'évaluer à la lumière des exigences de preuve convenues. Ils pourront ensuite :

- **Définir les exigences en matière de preuve** : Établir conjointement les preuves techniques précises, les données de référence et les indicateurs d'évaluation requis par les deux pays pour approbation, acquisition ou déploiement du système.
- **Mener une évaluation pilote** : Effectuer un essai au cours duquel les décideur-euse-s nationaux-ales évaluent les résultats fournis par le partenaire à la lumière de leurs propres exigences, en notant ce qui est accepté et ce qui nécessite une validation additionnelle.

2. Établir un registre de confiance transfrontalier.

Le projet pilote devrait mener à un rapport portant sur la manière dont chaque pays évalue et utilise les résultats d'évaluation de son partenaire. Le rapport devrait distinguer :

- **Preuves acceptées** : Les volets des évaluations provenant du partenaire qui peuvent être directement reconnus sans reprise des essais.
- **Preuves nécessitant une validation nationale** : Domaines demandant de nouveaux tests au niveau national, une validation sélective ou un accès sécurisé aux poids du modèle sous-jacent et aux données.
- **Évaluation nationale** : Exigences qui demandent une évaluation nationale indépendante parce que les conditions de déploiement diffèrent ou que les responsabilités nationales ne peuvent être déléguées.
- **Critères de réévaluation** : Des changements apportés au modèle ou à ses permissions, ou de nouvelles preuves de défaillance, qui nécessiteraient de réévaluer des conclusions précédemment acceptées.

Ce registre de confiance indiquerait ce dont chaque pays a besoin avant de se fier aux évaluations de l'autre. Il pourrait ensuite nourrir les décisions sur la possibilité d'élargir la coopération à d'autres applications liées à des infrastructures essentielles et, si approprié, des usages restreints de l'IA en matière de défense.