

Building Interoperable AI Assurance Between the UK and Canada

Authored by par

Connor Attridge, Research Associate - the Centre for Emerging
Technology and Security - The Alan Turing Institute

Henri Vilandré, Senior Policy Advisor - Mila

This commentary was developed in collaboration with The Alan Turing Institute's Centre for
Emerging Technology and Security ([CETaS](#)). A copy is also available on [their website](#).

Disclaimer : The opinions and views and opinions expressed in this paper belong solely to the
authors and do not represent the official policy or position the publishing institutes.

October 5th, 2026

Building Interoperable AI Assurance Between the UK and Canada

As more AI systems move into government, critical infrastructure and security settings, assurance can help organisations meet regulatory requirements and deploy AI safely. [Recent incidents](#) have shown that advanced AI agents can act beyond their intended scope – even during controlled evaluations. This emphasises the need for evaluation and assurance of the underlying models, but also of the operating conditions and institutional processes surrounding a system's use.

Canadian Prime Minister Mark Carney's September 2026 European visit underlined his [ambition](#) to deepen cooperation with like-minded partners on AI. In his address to the European Parliament, he proposed joint work on AI safety, standards and transparency. During the same visit, the UK and Canada agreed to [deepen cooperation](#) on safe and responsible AI. Canada and the UK, as Five Eyes allies with established information-sharing relationships, have a strong foundation for doing so. Working towards interoperable AI assurance – where assurance evidence can be shared, understood and used across jurisdictions – offers a way to turn this political ambition into practice.

The political basis for cooperation already exists. In June 2025, the UK and Canada [committed](#) to co-develop AI policy and standards, and to deepen collaboration on frontier systems for national security. This builds on cooperation in [AI compute](#), [safety research](#) and [researcher exchange](#). The practical question is how to translate those commitments into decisions about whether, and under what conditions, partner evidence can satisfy domestic assurance requirements.

For middle powers, AI sovereignty depends in part on retaining a choice of AI technologies and suppliers. Cooperation on assurance can help both countries assess those options. We propose a practical approach to UK–Canada cooperation on AI assurance with three main objectives:

1. **Facilitate safe deployment**, by building an agreed process for AI adoption across critical and high-stakes sectors.
2. **Preserve and reinforce middle-power autonomy**, where both nations jointly shape the defining criteria for safe AI while retaining control over domestic deployment decisions.
3. **Build institutional trust through joint evaluation work**, establishing when cooperation can extend to more sensitive domains.

Drawing on civil aviation's experience, the UK and Canada should test when a partner's evaluation findings can satisfy domestic assurance requirements.

Assurance and institutions

AI assurance, in the broad sense, is how an organisation satisfies itself that an AI system does what it claims to do, [safely and lawfully](#), in its intended context. Beyond this broad definition, the vocabulary is contested. [CETaS research](#) on the UK's defence and security assurance market found that assurance “means different things to different people”, spanning narrow technical testing, compliance processes and more holistic definitions. Agreeing on working definitions of assurance, verification and validation would help UK and Canadian evaluators clarify what partner testing establishes and what further evidence is needed to justify domestic deployment.

Two elements of assurance are relevant. First, assurance applies to a system in a specified deployment context. The evidence required depends on the system, its purpose, its operating conditions and the consequences of failure. Second, the methods for assuring these systems are themselves evolving. Benchmarking tests leak into training data, sandboxes can be too restrictive to be realistic or too permissive to be safe, and different red-teaming approaches find different classes of failure.

The measurement methods underpinning assurance are nonetheless becoming more structured. In January 2026, the US National Institute of Standards and Technology (NIST) published [draft practices](#) for automated benchmark evaluations of language models and AI agents. The International Network for Advanced AI Measurement, Evaluation and Science, of which Canada and the UK are members, has meanwhile developed [consensus practices](#) for automated evaluations and is building [common approaches](#) to third-party testing. Canada and the UK have already participated in [joint network work](#) on agentic evaluations, while the UK's AI Security Institute's (AISI) open-source [Inspect framework](#) and [engineering guidance](#) make evaluation tools reusable across organisations.

This demonstrates a basis for cooperation, but not a case for replacing different national evaluators with a single shared body. Reliance on one evaluator or methodology risks reproducing the same blind spots across systems. Evaluations conducted by different institutions can instead provide complementary evidence, particularly when those institutions share failed tests and methodological limitations as well as successful results. Interoperable assurance could therefore improve evaluation quality and coverage, as well as reduce duplicated testing or compliance costs.

Why the UK and Canada?

The UK and Canada are particularly well placed to test this because they share both a strategic middle-power position and complementary capabilities. Among AI middle powers, the UK has developed an established AI evaluation capability through AISI. Canada's newer evaluation system spans several institutions and is receiving further government investment. A new Canadian AI strategy, [AI for All](#), for example, commits a further C\$50 million to expand the Canadian AI Safety Institute's (CAISI) capacity to track emerging risks and conduct transparent evaluations.

The rationale for cooperation is not just built on allyship or the notion that their institutions are direct equivalents. Both countries share a [strategic interest](#) in retaining national control over how AI systems, developed within a largely US- and China-led technological ecosystem, are evaluated and deployed. Their partnership offers a practical test of whether middle powers can draw on one another's strengths without surrendering national responsibility for approval, procurement or deployment decisions.

Where can we start?

Not every AI use case warrants a bilateral mechanism. For some low-risk commercial uses, a formal process for sharing and reviewing evidence would cost more than its benefits justify. Defence is a stated priority for UK-Canada cooperation: their June 2025 [joint statement](#) frames frontier AI collaboration in national security terms, and both countries have close defence and intelligence relationships through alliances like Five Eyes. Some defence applications nevertheless introduce classification and operational constraints that could complicate an initial pilot.

Critical national infrastructure (CNI) offers several promising starting points. It has applications with direct defence and security implications, and established operating and regulatory requirements against which evidence can be assessed. Both countries are also strengthening cybersecurity oversight: Canada's [Critical Cyber Systems Protection Act](#) received Royal Assent in June 2026, with implementation being phased in, while the UK's [Cyber Security and Resilience Bill](#) is before the House of Lords. A pilot could build on this sectoral oversight, working with relevant operators and regulators to test how partner evidence can support decisions about AI deployment.

Risks vary across CNI applications and sectors. For example, an electricity-grid forecasting model that advises operators presents a different assurance problem from a water-treatment

controller that acts on its own. Both governments could select a bounded application where operating conditions, failure modes and human oversight are sufficiently understood to test cross-border reliance. The pilot should focus on a use case where assurance demands justify joint work and the evidence needed can be shared and assessed. This can be done without starting with the most sensitive and high-risk use cases.

Lessons from aviation

Cross-border assurance is not new or specific to AI. Aviation is one of several fields to work through comparable problems while balancing national authority and certification. Bilateral aviation safety agreements allow one authority to rely on specified technical findings produced by a trusted counterpart, within an agreed scope, while retaining its own regulatory responsibility.

The UK and Canada already operate this way. A [working arrangement](#) between the UK Civil Aviation Authority and Transport Canada has existed since 1 January 2021, and provides for reciprocal acceptance of airworthiness and maintenance findings. In June 2026, the two regulators [extended it](#) to mutual acceptance of recurrent flight-simulator evaluations.

Separately, the US Federal Aviation Administration's (FAA) safety risk matrix calibrates scrutiny to risk, grading potential failures by severity and probability into low (acceptable), moderate (acceptable with mitigation) and high (unacceptable) cases.

AI assurance must examine both the model and the conditions in which it is used. Uncertain or adversarial environments make that assessment harder. Recent [CETaS research](#) on behavioural assurance of agentic AI emphasises agents' actions and interactions, and the effectiveness of oversight. Shared evaluation reports should therefore describe the testing methods, the tools available to the system, its permissions and the organisational controls in place. This would help domestic assessors judge whether the findings apply to their intended use.

Table 1. FAA Aviation Safety Risk Matrix

Risk Probability	Risk Severity				
	Catastrophic (A)	Hazardous (B)	Major (C)	Minor (D)	Negligible (E)
Frequent (5)	5A	5B	5C	5D	5E
Occasional (4)	4A	4B	4C	4D	4E
Remote (3)	3A	3B	3C	3D	3E
Improbable (2)	2A	2B	2C	2D	2E
Extremely improbable (1)	1A	1B	1C	1D	1E

Low (acceptable)	Moderate (acceptable with mitigation)	High (unacceptable)
------------------	---------------------------------------	---------------------

Source: Adapted from FAA, “Safety Management Systems for Aviation Service Providers,” Advisory Circular, 21 May 2024.

A traffic-light model of reliance

The aviation safety approach can also be applied to AI assurance, using three tiers of reliance on partner assurance. The question is which domestic assurance requirements partner evidence can meet. We propose applying this approach first to bounded critical infrastructure applications:

1. **Green (greater reliance).** Bounded systems in controlled environments, such as a predictive maintenance system with human review and existing safety checks. Partner evaluations can be accepted with limited repetition and jointly agreed reporting, where they cover the intended domestic use.
2. **Yellow (conditional reliance).** Systems operating within defined limits but exposed to variable conditions, such as trading algorithms with exposure limits or supervised maritime surveying. Partner evidence informs domestic approval, supplemented by

selective retesting. Secure access to relevant model weights, data or other technical evidence may be needed.

3. **Red (national evaluation required).** Applications where failure could have severe consequences or national responsibilities require independent evaluation, such as AI supporting national electricity-grid restoration during a cyberattack. Partner evidence remains informative, but independent national evaluation and approval are required.

Table 2. Illustrative tiers of reliance on partner assurance

Operating environment	System complexity		
	Integrated/adaptive	Connected	Bounded
Contested/ adversarial	AI supporting national electricity-grid restoration during an ongoing cyberattack		
Variable		Supervised autonomous maritime surveying	
Controlled/ predictable		Trading algorithm with predefined exposure limits	AI-assisted predictive maintenance system
Greater reliance		Conditional reliance	National evaluation required

Note: Placements are illustrative and depend on deployment, safeguards and consequences of failure. Unlike aviation's risk categories, red denotes insufficient grounds for reliance on partner assurance alone. National decision-making authority is retained throughout.

Source: Authors' analysis.

A pilot should test these distinctions in practice: which findings can be accepted, which require domestic validation, and which judgements must remain national. Experience in lower-risk applications can build institutional trust and establish the conditions for wider cooperation.

Next steps

Two steps can be taken to test cross-border reliance:

1. Identify a joint critical infrastructure pilot.

The two governments should select one bounded, lower-sensitivity critical infrastructure use case and assess it against agreed evidence requirements. They would then:

- **Define evidence requirements:** Jointly establish the exact technical evidence, benchmark data and evaluation metrics required by both nations to approve, procure or deploy the system.
- **Run a trial evaluation:** Conduct a trial in which domestic decision-makers assess partner-produced findings against their own requirements, recording what is accepted and what needs further validation.

2. Establish a cross-border reliance record.

The pilot should produce a record of how each country assesses and uses its partner's evaluation findings. The record should distinguish:

- **Accepted evidence:** Components of partner evaluations that can be accepted directly without repetition.
- **Evidence requiring domestic validation:** Areas requiring domestic re-testing, selective validation, or secure access to underlying model weights and data.
- **National assessment:** Requirements that need independent domestic assessment because deployment conditions differ or national responsibilities cannot be delegated.
- **Reassessment triggers:** Changes to the model or its permissions, or new evidence of failure, that would require previously accepted findings to be reviewed.

That reliance record would show what each country needs before it can rely on the other's evaluations. It could then inform decisions about extending cooperation to additional critical infrastructure applications and, where appropriate, more bounded uses of AI in defence. Both countries could then draw more confidently on each other's evaluations while retaining responsibility for how AI is used.