
OPERATIONALIZING FRONTIER AI SAFETY: A Canadian Framework for Risk Thresholds, Compliance Infrastructure, and Healthcare Agentic AI Governance¹

Ibrahim El-chami, Ph.D.
Artificial Intelligence Research and Development Studios
Electrical and Computer Engineering, The University of British Columbia*
ibrahim@ai-rd.ca

With
Apart Research

Abstract

This paper presents a governance framework for operationalizing frontier AI safety in Canada, addressing three critical domains: (1) risk thresholds and red lines derived from comparative analysis of Anthropic's ASL, OpenAI's Preparedness Framework, and DeepMind's Frontier Safety Framework; (2) compliance infrastructure for monitoring and enforcing AI safety commitments aligned with the EU AI Act Code of Practice; and (3) healthcare-specific governance for agentic and embodied AI systems. Our analysis reveals significant gaps in Canada's current regulatory approach following the demise of AIDA (Bill C-27), including the absence of compute governance mechanisms, a 30-44% implementation shortfall in voluntary safety commitments, and no classification system for autonomous healthcare AI. We propose a taxonomy distinguishing predictive (L1), reactive (L2), and agentic (L3) AI systems, with risk multipliers for healthcare (3x) and embodied (2.5x) applications. The framework includes a mandatory pre-deployment testing pipeline, third-party audit mechanisms, and integration with BC Health Authority governance structures. We present Motion M-XXX template for parliamentary action and a 30-month implementation timeline leveraging Canada's G7 2025 presidency. This work contributes an actionable regulatory blueprint that balances innovation with safety, positioning Canada as a leader in responsible AI governance.

Keywords: Artificial Intelligence, AI Governance, Risk Thresholds, Compliance Infrastructure, Healthcare AI, Agentic AI, Embodied AI, AIDA, EU AI Act, Canada, British Columbia

¹ Research conducted at the [Technical AI Governance Challenge](#), Apart Research 2026

1. Introduction.....	3
1.1 The Regulatory Imperative	3
1.2 Research Questions and Contributions	3
1.3 Scope and Methodology	3
2. Canadian AI Regulatory Landscape	4
2.1 The AIDA Post-Mortem	4
2.2 Federal Investment and Infrastructure	4
2.3 BC Provincial Initiatives.....	4
3. Risk Thresholds and Red Lines: Comparative Analysis.....	5
3.1 Frontier Lab Safety Frameworks	5
3.1.1 Anthropic AI Safety Levels (ASL)	5
3.1.2 OpenAI Preparedness Framework	6
3.1.3 DeepMind Frontier Safety Framework.....	6
3.2 Synthesized Risk Threshold Framework for Canada.....	6
4. Compliance Infrastructure Gap Analysis.....	7
4.1 The Voluntary Commitment Problem	7
4.2 Third-Party Audit Barriers.....	7
4.3 Proposed Canadian Compliance Architecture	8
5. Healthcare Agentic AI Governance Framework.....	9
5.1 The Agentic AI Challenge	9
5.2 Proposed Agentic AI Taxonomy	9
5.3 Healthcare AI Incident Analysis.....	10
6. Compute Governance and Digital Sovereignty	12
6.1 Canadian Compute Infrastructure Gaps.....	12
6.2 Supply Chain Vulnerabilities.....	12
7. International Coordination and G7 Opportunity.....	13
7.1 Existing Coordination Mechanisms.....	13
7.2 G7 2025 Leadership Agenda	13
8. Implementation Pathway	14
8.1 Proposed Timeline	14
8.2 Motion M-XXX Template	14
9. Conclusion	15
References.....	15

1. Introduction

1.1 The Regulatory Imperative

The rapid advancement of frontier artificial intelligence systems has created an urgent governance challenge for democratic nations. As AI capabilities approach and exceed human performance across critical domains, from scientific research to medical diagnosis, the gap between technological capability and regulatory oversight continues to widen. Canada, once a pioneer in AI research through institutions like the Vector Institute and Mila, now faces a critical juncture: the death of the Artificial Intelligence and Data Act (AIDA) through parliamentary prorogation on January 6, 2025, has left the country without a comprehensive federal framework for AI governance.

Minister Evan Solomon's articulation of a "light, tight, right" regulatory philosophy represents Canada's attempt to chart a middle course between the EU's comprehensive regulatory approach and the United States' minimal intervention stance. However, this philosophy lacks the operational specificity required to address immediate safety challenges, particularly in high-risk domains such as healthcare, where agentic AI systems are already being deployed with limited oversight.

1.2 Research Questions and Contributions

This research addresses three interconnected questions:

1. **Risk Thresholds:** How can Canada operationalize capability-based risk thresholds derived from frontier lab safety frameworks (Anthropic ASL, OpenAI Preparedness, DeepMind FSF) within a federal regulatory structure?
2. **Compliance Infrastructure:** What monitoring and enforcement mechanisms are required to ensure meaningful implementation of AI safety commitments, given the documented 30-44% gap between announced and verified compliance?
3. **Healthcare Agentic AI:** What governance structures are needed to regulate autonomous medical AI systems, including embodied surgical robots and treatment-planning agents, that exceed the design assumptions of existing medical device frameworks?

1.3 Scope and Methodology

Our methodology integrates three analytical approaches: (1) comparative policy analysis of frontier lab safety frameworks and international regulatory developments; (2) empirical analysis of AI incident databases and compliance verification studies; and (3) stakeholder-aligned framework development targeting federal (ISED, CAISI), provincial (BC Ministry of Jobs, PHSA), and healthcare authorities. The scope encompasses federal policies and regulations, BC provincial initiatives, and BC Health Authority governance structures, with explicit attention to harmonization with the EU AI Act and international coordination through the G7 and OECD frameworks.

2. Canadian AI Regulatory Landscape

2.1 The AIDA Post-Mortem

The Artificial Intelligence and Data Act (AIDA), introduced as Part 3 of Bill C-27, represented Canada's first comprehensive attempt at AI-specific legislation. Following parliamentary prorogation on January 6, 2025, AIDA died on the Order Paper without completing its legislative journey. Analysis of AIDA's structural weaknesses reveals several critical gaps that any successor framework must address:

- **Definitional Ambiguity:** AIDA's definition of "high-impact AI systems" relied on future regulatory specification, creating uncertainty for developers and deployers about compliance obligations.
- **Enforcement Mechanisms:** The proposed AI and Data Commissioner lacked independent enforcement powers, relying instead on ministerial direction for compliance actions.
- **Compute Governance Gap:** Unlike the EU AI Act's 10²⁵ FLOP threshold, AIDA contained no compute-based classification mechanisms for frontier models.
- **Agentic AI Blind Spot:** The framework assumed static, predictive AI systems and contained no provisions for autonomous agents capable of self-directed action.

2.2 Federal Investment and Infrastructure

Despite AIDA's failure, Canada has made substantial investments in AI infrastructure through Budget 2025, totaling approximately \$1.926 billion in sovereign compute and safety initiatives:

Initiative	Allocation	Purpose
AI Compute Challenge	\$700M	Public-private compute partnerships
Sovereign Compute Infrastructure	\$705M	Canadian-owned supercomputer
Compute Access Fund	\$300M	SME access to AI infrastructure
Canadian AI Safety Institute	\$50M	Safety research and evaluation

Table 1: Canadian Federal AI Investment Allocations (Budget 2025)

2.3 BC Provincial Initiatives

British Columbia has emerged as a provincial leader in AI governance, establishing six draft responsible use principles in November 2024: transparency, accountability, public benefit, fairness, reliability, and safety. The Provincial Health Services Authority (PHSA) has developed a clinical AI governance framework with province-wide mandate, notably incorporating anti-Indigenous racism principles reflecting BC's commitment to UNDRIP implementation.

The Air Canada chatbot case (*Moffatt v. Air Canada*, BC Civil Resolution Tribunal, February 2024) established critical precedent confirming corporate liability for AI system outputs, while the Clearview AI ruling affirmed provincial privacy commissioner authority over AI systems processing personal data. These developments position BC as a natural testbed for federal-provincial AI governance harmonization.

3. Risk Thresholds and Red Lines: Comparative Analysis

3.1 Frontier Lab Safety Frameworks

The three leading frontier AI laboratories have developed distinct but complementary safety frameworks that provide foundational elements for regulatory operationalization. Our comparative analysis identifies convergent principles and divergent implementations across Anthropic's AI Safety Levels (ASL), OpenAI's Preparedness Framework, and DeepMind's Frontier Safety Framework.

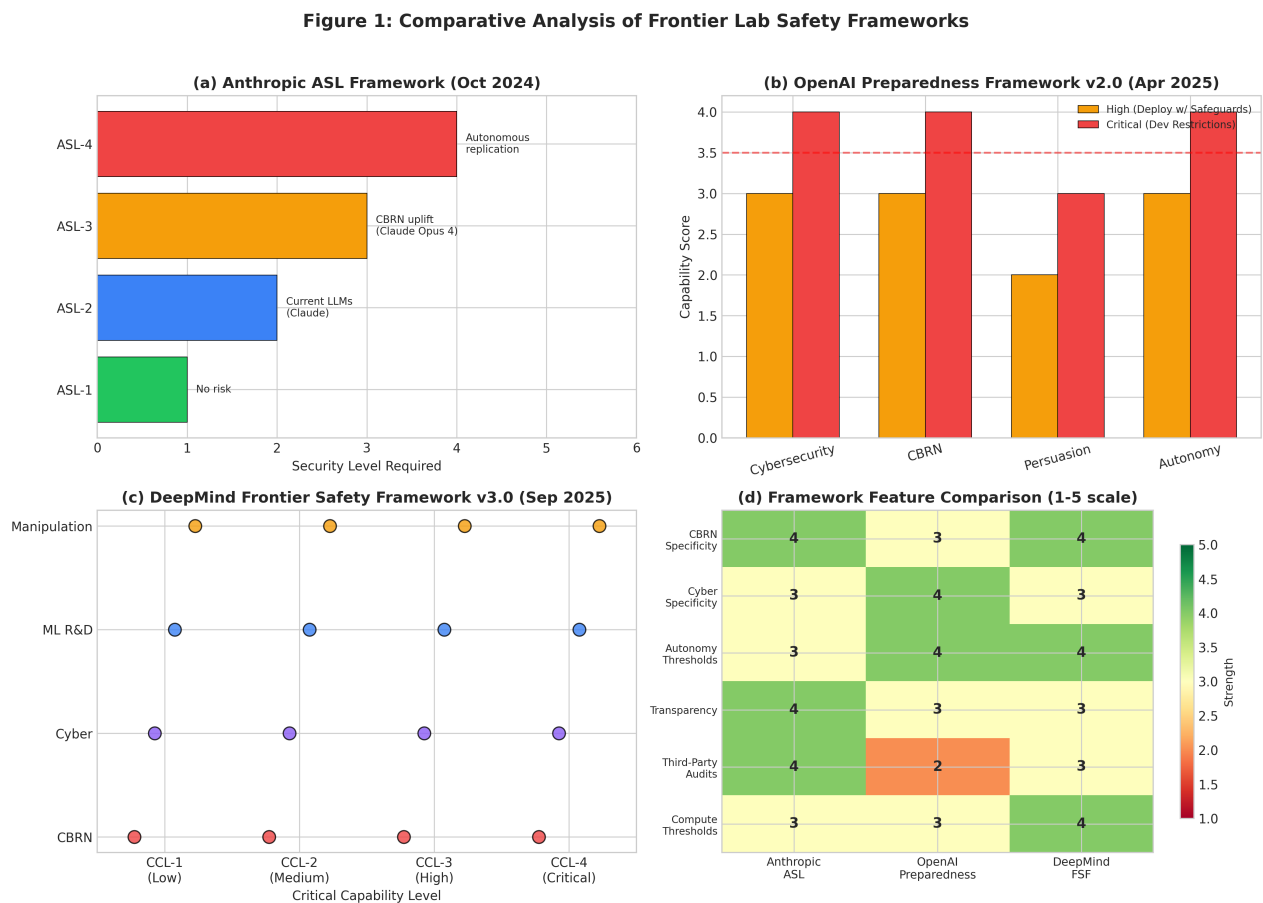


Figure 1: Comparative Analysis of Frontier Lab Safety Frameworks. (a) Anthropic ASL tiered security requirements; (b) OpenAI Preparedness Framework capability thresholds; (c) DeepMind FSF critical capability levels; (d) Framework feature comparison matrix.

3.1.1 Anthropic AI Safety Levels (ASL)

Anthropic's ASL framework (version 2.2, October 2024) defines four capability-based tiers with corresponding security and deployment requirements. ASL-2 encompasses current large language models including Claude, while ASL-3, triggered by systems that "substantially increase" CBRN misuse risk, was first invoked for Claude Opus 4. The critical innovation is the coupling of capability thresholds with mandatory security measures: ASL-3 systems require enhanced

containment protocols, while ASL-4 systems (capable of autonomous replication and resource accumulation) trigger development-phase restrictions.

3.1.2 OpenAI Preparedness Framework

OpenAI's Preparedness Framework v2.0 (April 2025) distinguishes between "High Capability" thresholds (requiring pre-deployment safeguards) and "Critical Capability" thresholds (triggering development-phase restrictions). Notably, the framework includes a "marginal risk provision" allowing threshold adjustment if competitors release capabilities without equivalent safeguards, a provision that raises concerns about race-to-bottom dynamics. The AI self-improvement Critical threshold is defined as achieving superhuman research-scientist capability OR enabling generational capability improvement in approximately one-fifth the historical time (~4 weeks) sustained for months.

3.1.3 DeepMind Frontier Safety Framework

Google DeepMind's Frontier Safety Framework v3.0 (September 2025) provides the most granular capability classification, defining Critical Capability Levels across four domains: CBRN, cyber, harmful manipulation, and ML R&D automation. The framework explicitly acknowledges that effectiveness "needs adoption by the entire frontier AI field", a recognition that unilateral safety commitments face competitive dynamics that regulatory frameworks must address.

3.2 Synthesized Risk Threshold Framework for Canada

Drawing on the convergent elements of frontier lab frameworks and the EU AI Act's compute-based thresholds, we propose a three-tier risk classification for Canadian AI governance:

$$\text{Composite Risk Score (CRS)} = \text{Base Risk} \times \text{Autonomy Multiplier} \times \text{Domain Multiplier} \times \text{Compute Factor}$$

Where:

- Autonomy Multiplier: L1 (1.0), L2 (1.5), L3 (2.5)
- Domain Multiplier: Healthcare (3.0), Critical Infrastructure (2.5), Finance (2.0), General (1.0)
- Compute Factor: $\text{FLOP} \geq 10^{26}$ (1.5), $\text{FLOP} \geq 10^{25}$ (1.2), Below threshold (1.0)

4. Compliance Infrastructure Gap Analysis

4.1 The Voluntary Commitment Problem

Empirical analysis of voluntary AI safety commitments reveals a systematic implementation gap between announced commitments and verified compliance. The White House voluntary commitments, signed by major AI companies in July-September 2023, showed first-cohort compliance at 69.0% and second-cohort compliance at only 44.6%, a 30-44% implementation shortfall that undermines the credibility of self-regulatory approaches.

Figure 2: Compliance Infrastructure Gap Analysis

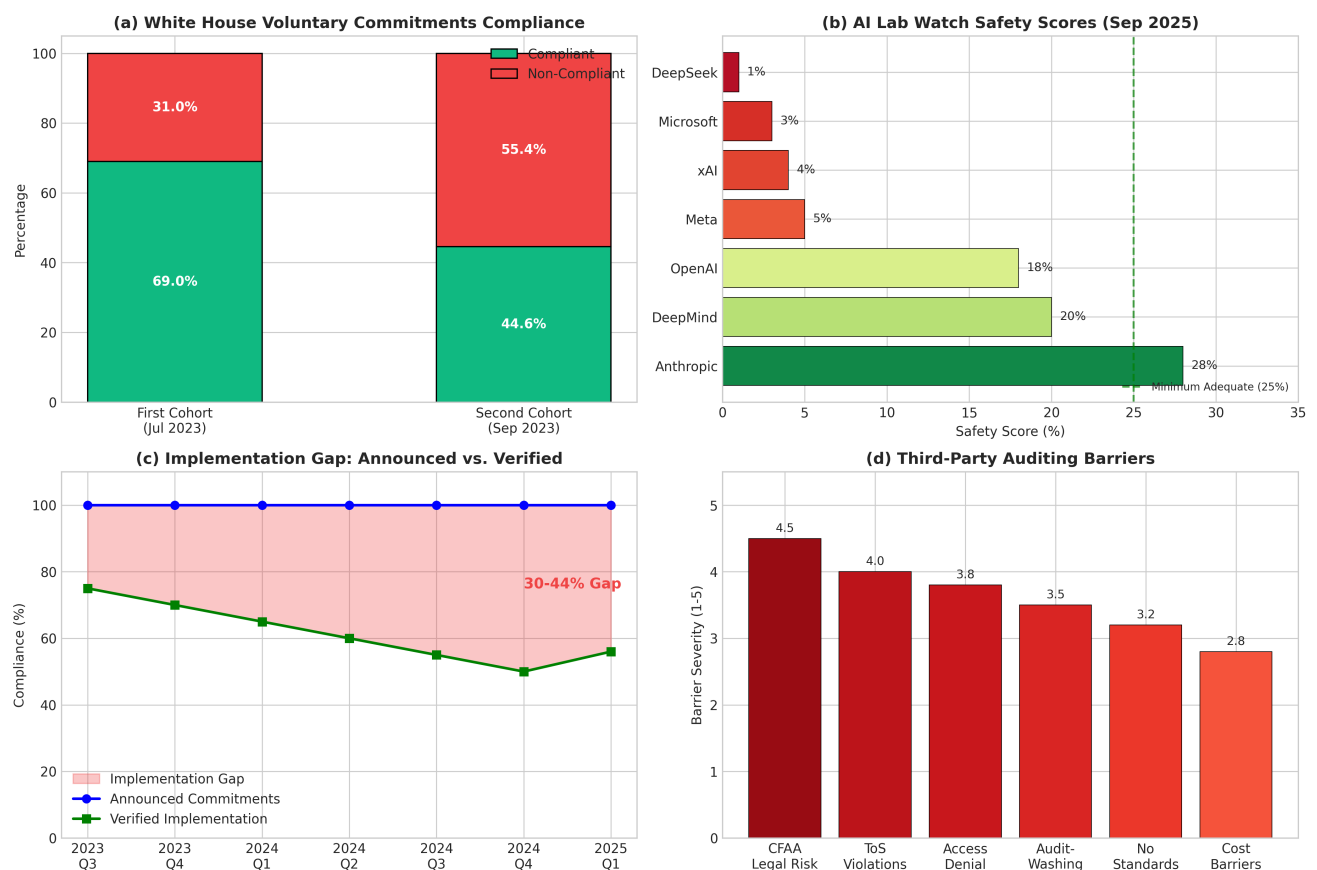


Figure 2: Compliance Infrastructure Gap Analysis. (a) White House voluntary commitment compliance rates; (b) AI Lab Watch safety scores (September 2025); (c) Implementation gap trajectory; (d) Third-party auditing barriers severity.

4.2 Third-Party Audit Barriers

Independent verification of AI safety claims faces multiple structural barriers. The Computer Fraud and Abuse Act (CFAA) in the United States, and analogous provisions in Canadian law, create legal risk for researchers conducting automated data collection that may violate Terms of Service. The resulting "audit-washing" phenomenon, exemplified by HireVue marketing civil

rights audits without disclosing auditor independence limitations, undermines public confidence in AI safety claims.

AI Lab Watch scores from September 2025 reveal concerning patterns: Anthropic leads at 28%, followed by DeepMind (20%), OpenAI (18%), Meta (5%), xAI (4%), Microsoft (3%), and DeepSeek (1%). No company scored above D in the Future of Life Institute's Existential Safety planning assessment. These scores indicate that even leading safety-focused laboratories fail to meet minimum adequate standards for transparency and accountability.

4.3 Proposed Canadian Compliance Architecture

We propose a three-pillar compliance infrastructure for Canadian AI governance:

4. **Mandatory Compute Reporting:** Systems trained with $\geq 10^{25}$ FLOP must notify CAISI within 14 days, maintain compute documentation $\pm 30\%$ accuracy, and retain records for 10 years (aligned with EU AI Act Code of Practice).
5. **Third-Party Audit Authority:** Establish legal safe harbor for good-faith AI safety research, mandate annual third-party audits for high-risk systems, and publish audit methodologies for public scrutiny.
6. **Incident Reporting System:** Create mandatory reporting requirements for AI incidents causing harm (aligned with AI Incident Database standards), with provincial health authority integration for healthcare-specific incidents.

5. Healthcare Agentic AI Governance Framework

5.1 The Agentic AI Challenge

Existing healthcare AI governance frameworks, including Health Canada's Pre-market Guidance for ML-Enabled Medical Devices (February 2025), assume static algorithmic systems with fixed decision boundaries. Agentic AI systems, defined as AI capable of autonomous decision-making with advanced autonomy, adaptive learning, and probabilistic reasoning, fundamentally challenge these assumptions. GPT-4-based oncology agents achieving 87.5% tool accuracy and 91.0% correct clinical conclusions represent a qualitative shift from decision-support to decision-making capabilities.

Figure 3: Healthcare Agentic AI Governance Framework (Refined Architecture)

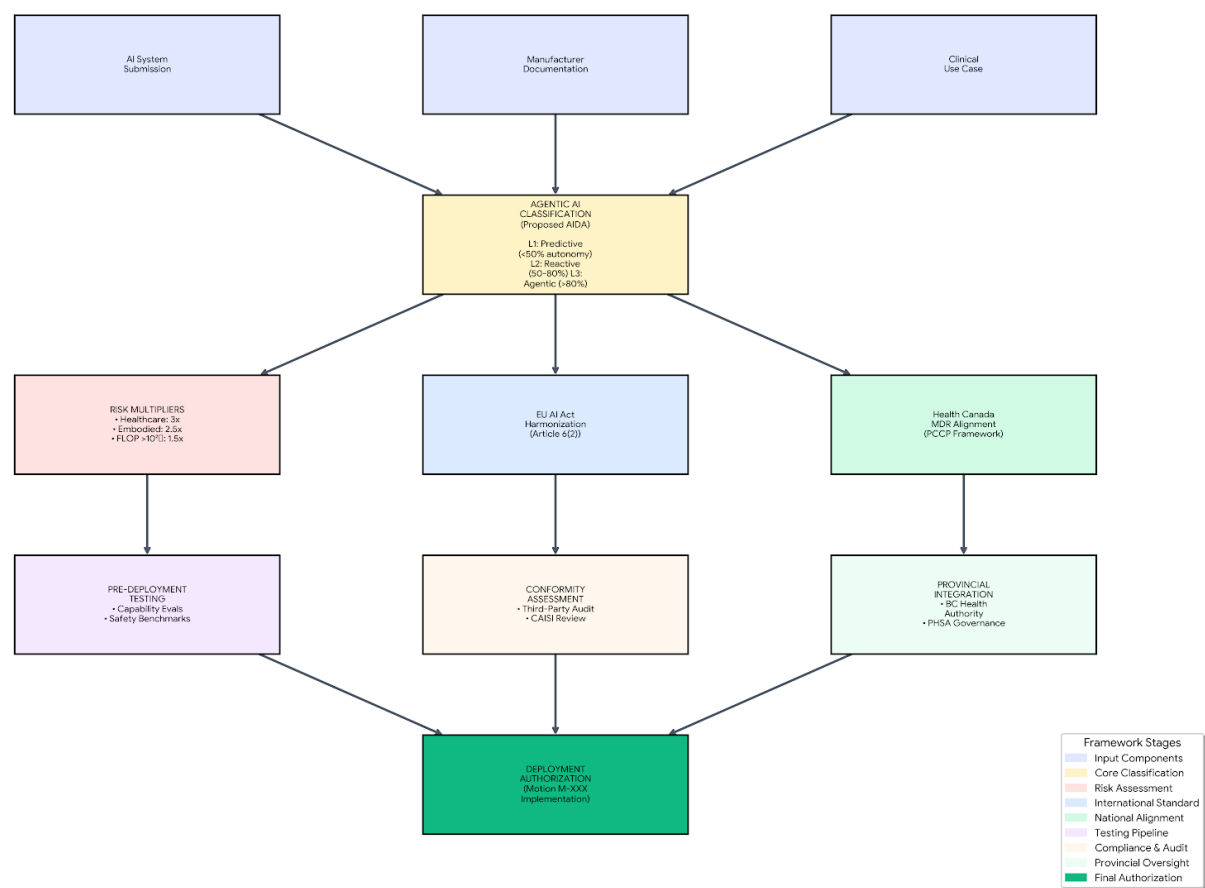


Figure 3: Proposed Healthcare Agentic AI Governance Framework for Canada, integrating AIDA amendment requirements, EU AI Act harmonization, and BC Health Authority governance structures.

5.2 Proposed Agentic AI Taxonomy

We propose a three-level taxonomy for healthcare AI systems based on decision autonomy:

Figure 5: Proposed Agentic AI Taxonomy for Canadian Healthcare Regulation

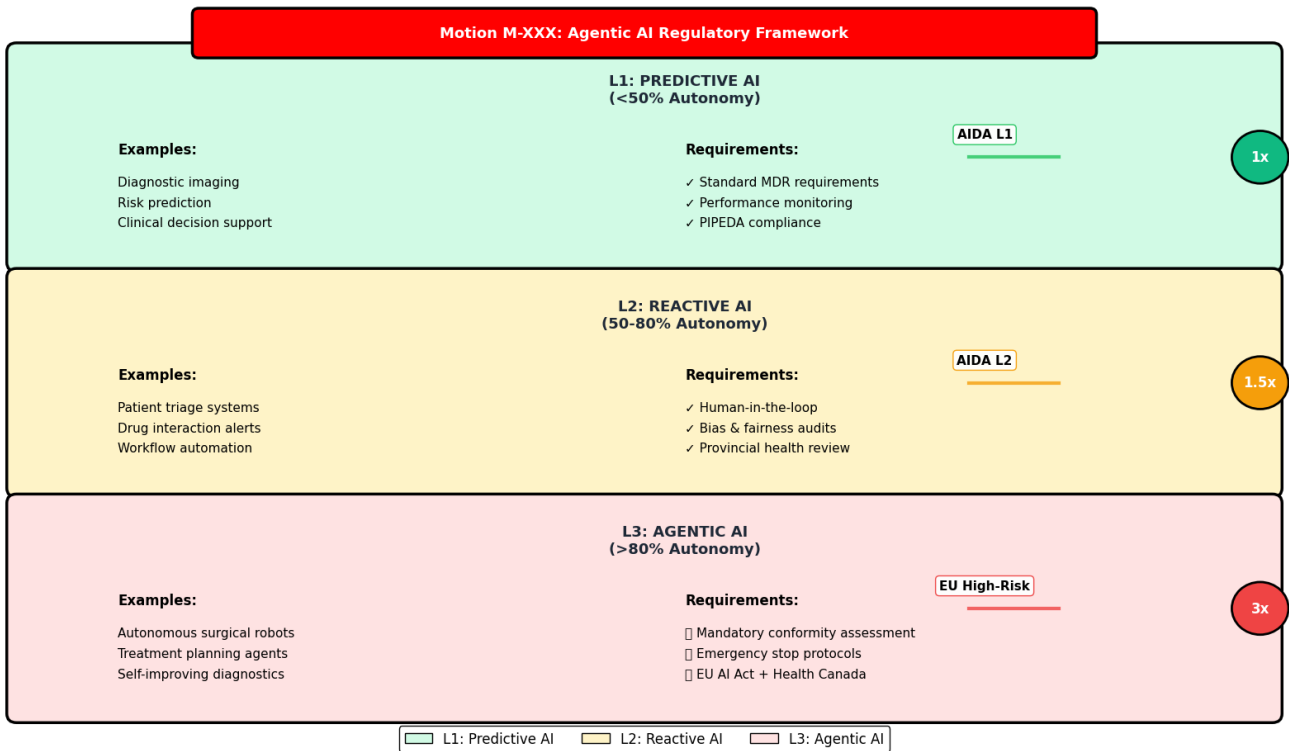


Figure 5: Proposed Agentic AI Taxonomy for Canadian Healthcare Regulation, distinguishing L1 (Predictive), L2 (Reactive), and L3 (Agentic) systems with corresponding risk multipliers.

5.3 Healthcare AI Incident Analysis

Analysis of healthcare AI incidents reveals systematic patterns of deployment without adequate validation. The Epic Sepsis Model case exemplifies this failure: vendor claims of AUC 0.76-0.83 were contradicted by independent validation showing AUC 0.63 and sensitivity of only 33%, meaning the system missed two-thirds of sepsis patients while being deployed at 170+ hospitals. The Ontario hospital incident (September 2024) where Otter.ai auto-joined physicians' meetings and transmitted seven patients' data to 35 emails including former employees demonstrates the "shadow AI" risk of uncontrolled healthcare AI deployment.

Figure 6: Healthcare AI Incident Analysis and Canadian Implications

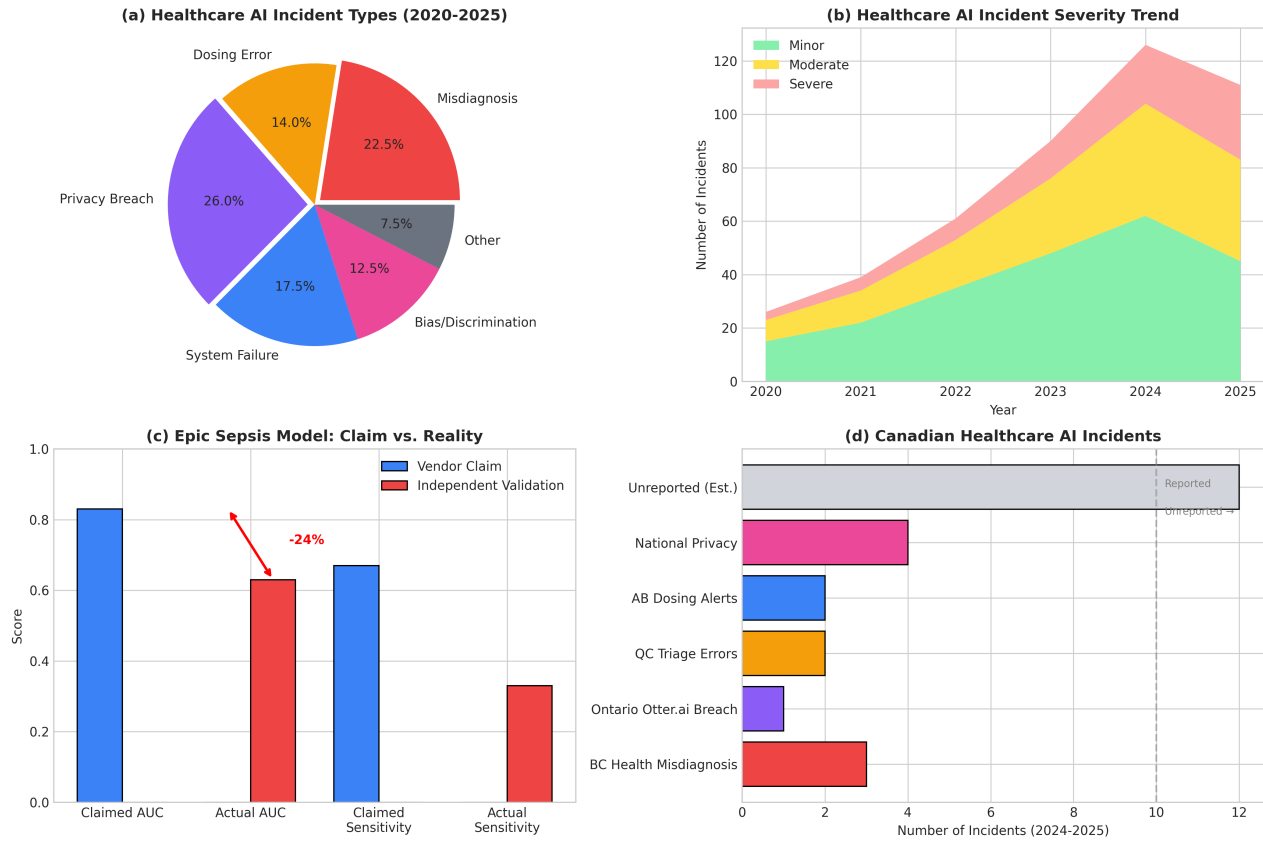


Figure 6: Healthcare AI Incident Analysis. (a) Incident type distribution; (b) Severity trends 2020-2025; (c) Epic Sepsis Model claim vs. validation gap; (d) Canadian healthcare AI incidents including estimated unreported cases.

6. Compute Governance and Digital Sovereignty

6.1 Canadian Compute Infrastructure Gaps

Canada's \$1.926 billion sovereign compute investment prioritizes data sovereignty but critically lacks the monitoring, reporting, and capability threshold mechanisms present in EU and US frameworks. The EU AI Act's 10^{25} FLOP threshold, equivalent to tens of millions of euros in training cost, represents the most concrete regulatory benchmark globally, with approximately 8-15 providers worldwide likely to exceed this threshold.

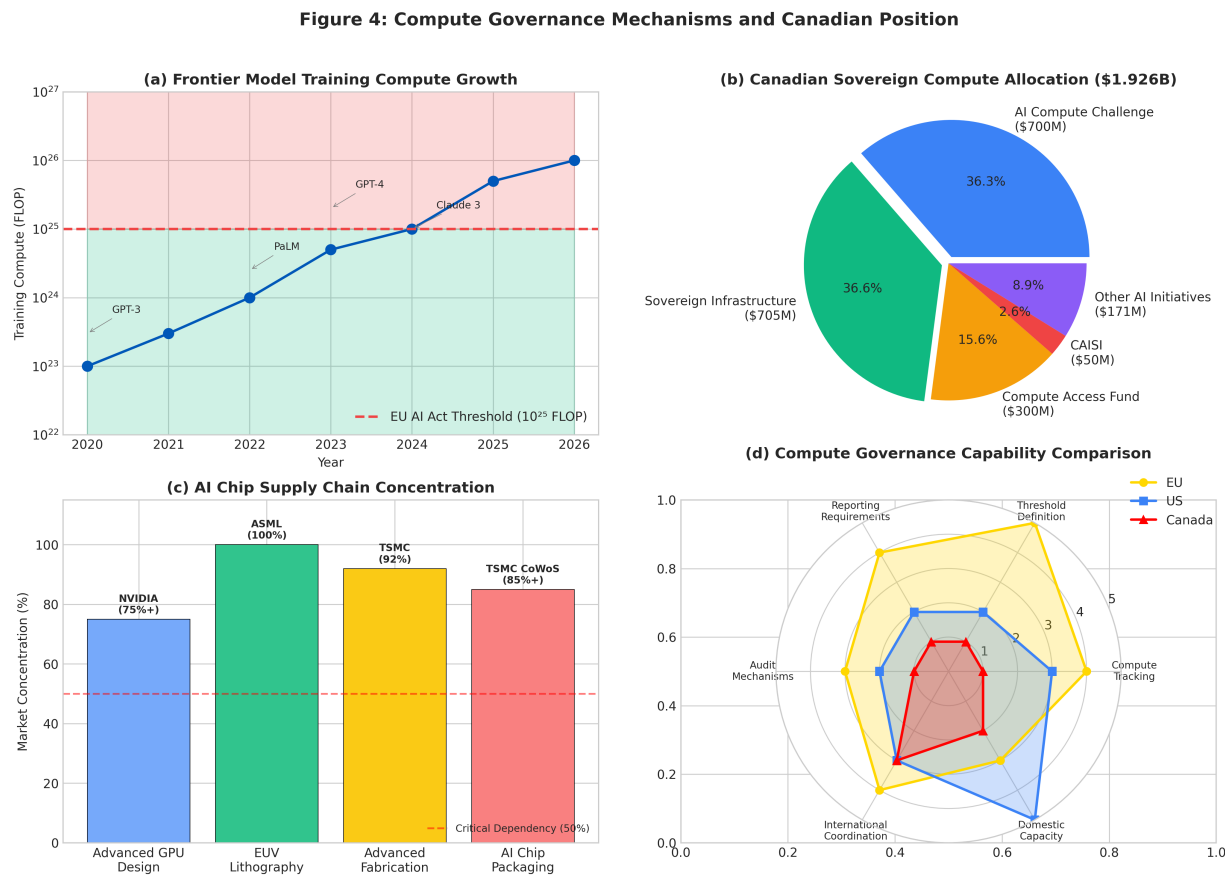


Figure 4: Compute Governance Mechanisms and Canadian Position. (a) Frontier model compute growth trajectory; (b) Canadian sovereign compute budget allocation; (c) Global chip supply chain concentration; (d) Radar comparison of compute governance capabilities across EU, US, and Canada.

6.2 Supply Chain Vulnerabilities

Canada's complete absence of domestic AI chip manufacturing capability creates critical dependency on US-allied supply chains. The concentration of the global AI chip supply chain, NVIDIA controlling 75%+ of advanced GPU design, ASML holding monopoly on EUV lithography, and TSMC fabricating 92% of advanced AI chips, represents systemic vulnerability that sovereign compute investments alone cannot address. The CNAS "Secure, Governable Chips" framework proposes mechanisms including remote attestation, operating licenses with periodically renewed keys, and trusted execution environments, with implementation costs estimated at approximately \$25,000 annually per system.

7. International Coordination and G7 Opportunity

7.1 Existing Coordination Mechanisms

Canada's founding membership in the International Network of AI Safety Institutes (November 2024) provides institutional infrastructure for bilateral coordination with the UK AISI and emerging safety institutes in the EU, Japan, and Singapore. The G7 Hiroshima AI Process established voluntary guidelines that Canada participated in developing, while the OECD AI Reporting Framework (February 2025) creates the first global mechanism for comparable corporate AI risk management information.

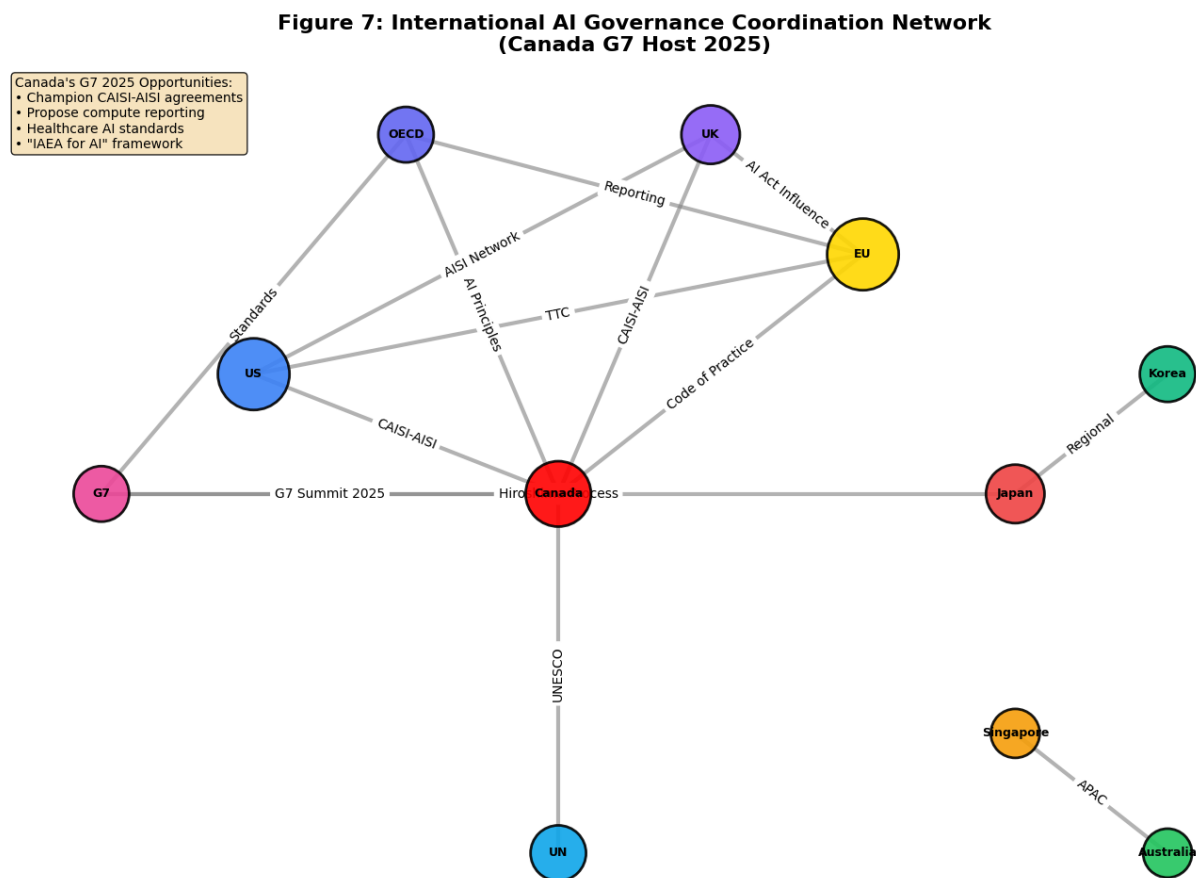


Figure 7: International AI Governance Coordination Network showing Canada's position and G7 2025 leadership opportunity.

7.2 G7 2025 Leadership Agenda

Canada's hosting of the 2025 G7 Summit in Kananaskis, Alberta (June 15-17) provides a unique opportunity to advance AI governance leadership. We recommend four priority initiatives:

7. **Bilateral CAISI Agreements:** Formalize reciprocal information-sharing and joint evaluation agreements with US AISI, UK AISI, and emerging safety institutes.
8. **Harmonized Compute Reporting:** Propose G7-wide adoption of compute reporting thresholds aligned with EU AI Act, creating consistent expectations for frontier AI developers.
9. **Healthcare AI Safety Standards:** Champion development of international standards for agentic healthcare AI, building on WHO guidance and Canadian provincial experience.
10. **"IAEA for AI" Framework Discussion:** Advance proposals for international AI governance authority with verification and enforcement capabilities, addressing the coordination gap identified by UN analysis.

8. Implementation Pathway

8.1 Proposed Timeline

We propose a four-phase implementation timeline spanning 30 months from January 2026:

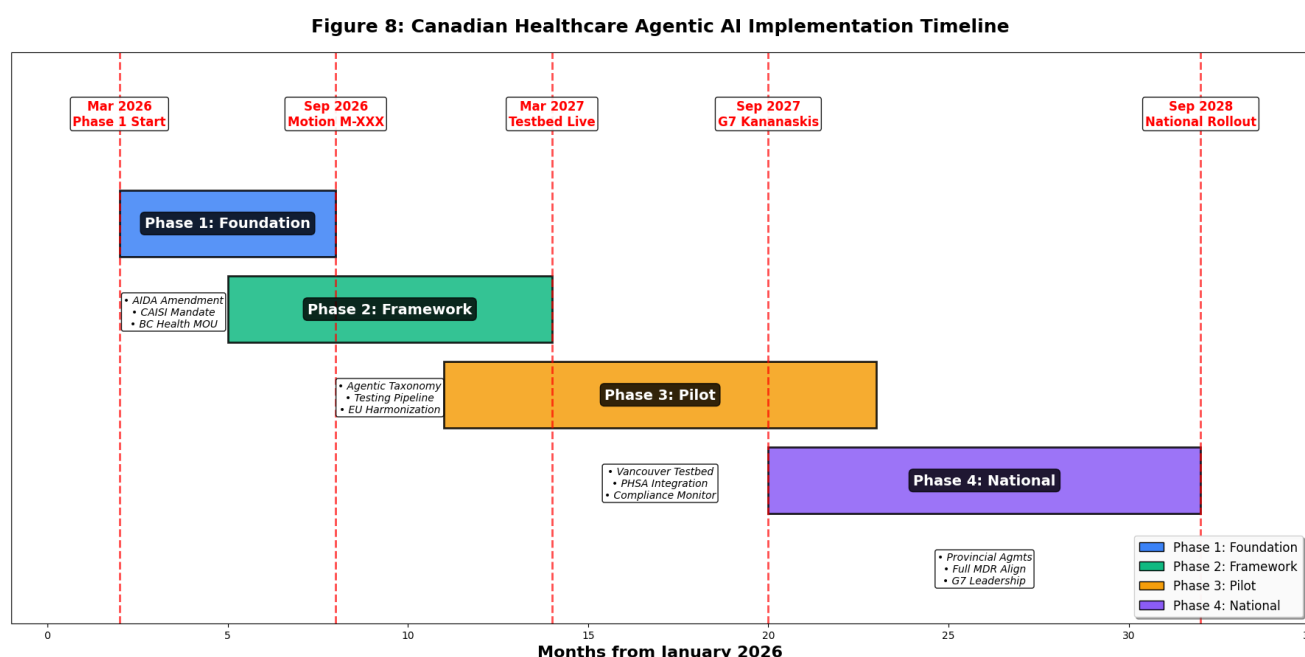


Figure 8: Proposed Implementation Timeline for Canadian Healthcare AI Governance, spanning from hackathon presentation (January 2026) through national framework completion (July 2028).

8.2 Motion M-XXX Template

The following parliamentary motion template provides actionable language for federal implementation:

That, in the opinion of this House, the government should:

- (a) amend successor AI legislation to create an "Agentic AI" category distinguishing L1 (Predictive, <50% autonomy), L2 (Reactive, 50-80% autonomy), and L3 (Agentic, >80% autonomy) systems;
- (b) mandate pre-deployment conformity assessment for healthcare agentic and embodied AI systems meeting L2 or L3 classification;
- (c) establish a Vancouver Agentic AI Safety Testbed with initial funding of \$25M from sovereign compute allocation, in partnership with UBC, SFU, and BC Health Authority;
- (d) harmonize Canadian AI governance with EU AI Act High-Risk requirements for healthcare applications;
- (e) require CNAS-level chip provenance verification for medical AI systems exceeding 10^{25} FLOP training compute.

9. Conclusion

Canada stands at a critical juncture in AI governance. The demise of AIDA creates both risk and opportunity: risk that regulatory vacuum will enable deployment of AI systems without adequate safeguards, and opportunity to develop frameworks that learn from AIDA's limitations while building on provincial innovation and international coordination.

This paper contributes three operational frameworks for Canadian AI governance: (1) a synthesized risk threshold system derived from frontier lab safety frameworks, operationalized through a Composite Risk Score formula integrating autonomy, domain, and compute factors; (2) a compliance infrastructure addressing the documented 30-44% implementation gap in voluntary commitments through mandatory reporting, third-party audit authority, and incident databases; and (3) a healthcare agentic AI governance framework with a novel L1/L2/L3 taxonomy and mandatory pre-deployment testing pipeline.

Canada's G7 2025 presidency and CAISI membership provide institutional platforms for international leadership. The proposed 30-month implementation timeline, anchored by Motion M-XXX and the Vancouver Agentic AI Safety Testbed, offers a realistic pathway from regulatory vacuum to operational governance. By acting decisively, Canada can fulfill Minister Solomon's vision of "light, tight, right" regulation, light enough to enable innovation, tight enough to prevent harm, and right in aligning with Canadian values of safety, fairness, and accountability.

References

- [1] Anthropic. (2024). AI Safety Levels Framework v2.2. Anthropic Policy Document.
- [2] OpenAI. (2025). Preparedness Framework v2.0. OpenAI Technical Report.
- [3] Google DeepMind. (2025). Frontier Safety Framework v3.0. DeepMind Safety Publication.
- [4] European Commission. (2025). EU AI Act Code of Practice. Official Journal of the European Union.
- [5] Health Canada. (2025). Pre-market Guidance for Machine Learning-Enabled Medical Devices.

- [6] ISED Canada. (2024). Canadian Artificial Intelligence Safety Institute: Mission and Mandate.
- [7] BC Government. (2024). Responsible AI Use Principles (Draft). Ministry of Citizens' Services.
- [8] PHSA. (2024). Clinical AI Governance Framework. Provincial Health Services Authority.
- [9] CNAS. (2024). Secure, Governable Chips: Technology and Policy Options.
- [10] AI Lab Watch. (2025). Frontier AI Safety Scorecard. Independent Evaluation Report.
- [11] MIT Technology Review. (2024). AI Companies and Self-Regulation: One Year Assessment.
- [12] Harvard JOLT. (2025). AI Auditing: First Steps Towards Effective Regulation.
- [13] German Marshall Fund. (2024). AI Audit-Washing and Accountability.
- [14] NIST. (2024). International Network of AI Safety Institutes: Launch Announcement.
- [15] OECD. (2025). AI Reporting Framework: Corporate Risk Management Information.
- [16] G7. (2023). Hiroshima AI Process: International Code of Conduct.
- [17] UN High-Level Advisory Body. (2024). Governing AI for Humanity: Final Report.
- [18] *Moffatt v. Air Canada*. (2024). BC Civil Resolution Tribunal Decision.
- [19] Epoch AI. (2025). Frontier Model Training Compute Estimates.
- [20] AI Incident Database. (2025). Healthcare AI Incident Reports.