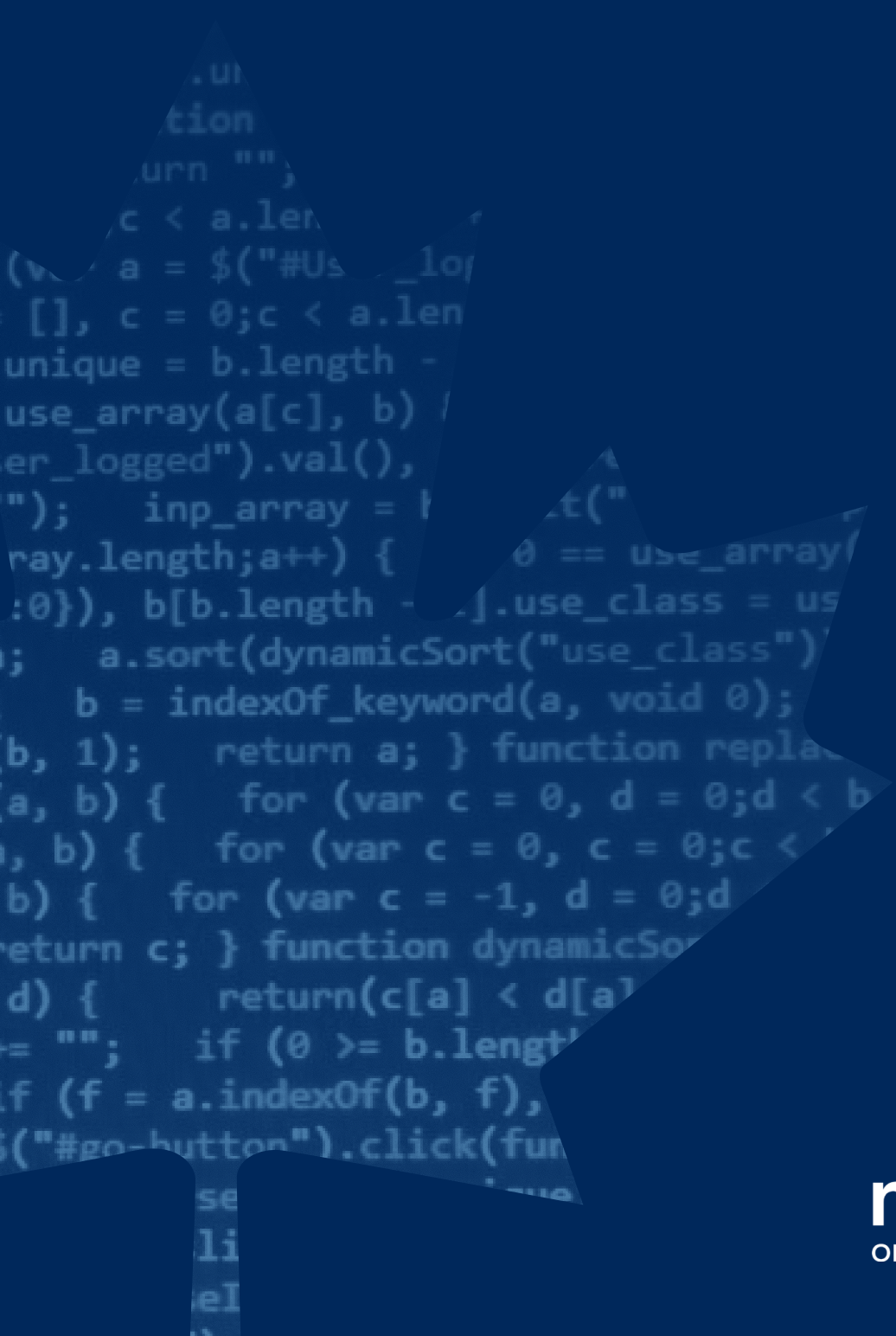
Sovereign by Design: Strategic Options for Canadian AI Sovereignty



March 2026

Sean Mullin and Jaxson Khan

munk school
OF GLOBAL AFFAIRS & PUBLIC POLICY

AUTHORS

Jaxson Khan, Senior Fellow

Jaxson Khan (www.jaxson.org) is CEO of Aperture AI, a strategic consultancy enabling major corporations and governments to lead on AI and digital technology. Jaxson is a Senior Fellow and Lecturer at the Munk School of Global Affairs and Public Policy and a Board Director of the Human Feedback Foundation.

Prior, Jaxson served as Senior Policy Advisor to Canada's Minister of Innovation, Science, and Industry, where he played a pivotal role in developing a $2.4 billion investment into AI infrastructure. His files included quantum, semiconductors, competition, intellectual property, and innovation policy. He also served as Chief of Staff at LawAI, an independent think tank.Before his AI policy work, he worked in several startups including as Director of Growth at Fable, an accessibility technology company.

Jaxson holds an EMBA from Quantic and a BA from Huron University, where he was recognized with an Alumni Award of Distinction and studied as a 3M National Student Fellow.

Sean Mullin, Senior Fellow

Sean Mullin (www.seanmullin.ca) is an economist and policy leader with over 20 years of experience at the intersection of technology, economic strategy, and public policy. He is currently a Senior Fellow at the Munk School of Global Affairs and Public Policy at the University of Toronto, where he focuses on AI policy and the economics of AI.

Sean previously served as Special Advisor to the Prime Minister for Economic Affairs, where he was the principal economic advisor to Prime Minister Trudeau and helped lead the design and launch of Canada's $2.4 billion Sovereign AI Compute Strategy. Prior to joining the PMO, Sean founded and led the Brookfield Institute for Innovation + Entrepreneurship, building it into Canada's leading innovation policy think tank, and served in senior advisory roles on economic and fiscal policy in the Ontario Premier's Office.

Sean holds an MBA from Oxford University, an MA in Economics from McGill, and an Honours BSc in Economics and Computer Science from the University of Toronto.

DOI: [PLACEHOLDER]
Munk School of Global Affairs & Public Policy
University of Toronto
315 Bloor Street West
Toronto, ON M5S 0A7
munkschool.utoronto.ca

munk school
OF GLOBAL AFFAIRS & PUBLIC POLICY

UNIVERSITY OF TORONTO

"The stakes are stack-level choices — black-box dependence or modular improvisation; opacity or legibility; someone else's roadmap or a sovereign design of your own. In practical terms, the decision is the difference […] between consuming intelligence as a service and composing it as an act of sovereignty. One rents a mind, the other trains its own in the wild."

– A Third Path For AI Beyond The US-China Binary, Dang Nguyen

ACKNOWLEDGEMENTS:

The authors thank the following contributors, whose help and advice were instrumental in developing this paper. Any errors or omissions are solely the responsibility of the authors.

Prof. Janice Stein – Project Advisor
Vice-Admiral (Ret'd) Ron Lloyd – Project Advisor
Iain Stewart – Project Advisor
Sydney Wisener – Research Assistant
Sabreena Shukul – Research Assistant
Dara Poizner – Copy Editor
Karen Gondard – Graphic Designer

The authors would also like to thank dozens of experts and practitioners from the public and private sectors who participated in interviews for this project.

This report was generously supported by funding from RBC, Cohere, Bell Canada, OpenText, and the Ontario Teachers' Pension Plan. The Munk School of Global Affairs and Public Policy maintained full editorial independence, and the funders did not influence the report's content or findings. The views expressed are those of the authors alone.

CONTENTS

EXECUTIVE SUMMARY

Artificial intelligence is advancing at an extraordinary speed, reshaping economies, national security, and geopolitical competition. Global AI investment is reaching historically unprecedented levels at hundreds of billions of dollars annually, and the United States now explicitly frames AI dominance as a core national security interest. Canada is the birthplace of modern AI, yet controls neither the major firms that dominate its deployment nor the critical infrastructure that powers it. This dependency is a dangerous vector for coercion, and recent tariff threats, territorial provocations, and "51st state" rhetoric have clarified that the rules-based order Canada once relied upon can no longer be assumed.

AI SOVEREIGNTY: This paper argues that sovereignty in the AI era means freedom from coercion, not digital isolationism or technological self-sufficiency. No country can achieve complete independence across the AI technology stack; the question is how to structure dependencies to preserve choice, reduce foreign leverage, and ensure that Canadian data and infrastructure remain governed by Canadian laws and values. We assess sovereignty across five dimensions (jurisdictional, operational, technological, societal, and economic) and find that AI amplifies threats across all five simultaneously (Section V).

GLOBAL CONTEXT: The United States pursues explicit technological dominance, with American platforms serving as vectors for American jurisdiction and power. China is building a self-sufficient AI stack while exporting open-weight models that embed its own political assumptions globally. The European Union combines regulation with industrial capacity-building. Middle powers face a choice between dependency on foreign AI systems or technological weakness — but coalition-building and hybrid strategies offer a path beyond this binary. Canada must navigate these dynamics, including the impending July 2026 CUSMA review, which presents both risk and opportunity for AI and digital sovereignty (Sections III, VI).

WINDOW FOR ACTION: The majority of new AI investment over the next half decade will reshape the technology stack, particularly for inference and deployment: the operational layer where AI systems process data, serve users, and generate value. Decisions being made today about infrastructure, platforms, and standards will shape the landscape for a generation, and making these decisions without regard for sovereignty risks locking Canada into dependencies that will be difficult to reverse. But the architecture is not yet settled, and a deliberate, sovereignty-oriented approach can influence how resources are deployed while the window for action remains open.

NAVIGATING TRADE-OFFS: Sovereignty does not come free. Sovereign infrastructure carries a cost premium, and the most secure options may constrain access to cutting-edge capabilities. Failing to adopt AI is itself a sovereignty risk: a country that falls behind becomes economically weaker, less competitive, and strategically vulnerable. This report presents a pragmatic, risk-based framework that calibrates measures to protect sovereignty with respect to their impact on economic competitiveness and the financial viability of the "sovereign solutions." The options presented are a menu for policymakers: different interventions can be pursued independently, at different speeds, and in different combinations.

CANADA'S STRENGTHS: Our layer-by-layer assessment of the AI technology stack (Section VII) reveals genuine strengths. Canada's clean energy advantage has attracted substantial data centre investment, and Canadian-owned operators provide some domestic alternatives to foreign hyperscalers at the cloud infrastructure level. Valuable government and private sector datasets in health, finance, law, and administration are strategic assets for sovereign AI development, particularly if Canada chooses to invest in protecting and developing them. Some Canadian AI application-layer companies have achieved global scale and multi-billion-dollar valuations. And a strong open-source ecosystem means sovereignty at the model operations layer derives primarily from infrastructure choices further down the stack.

CANADA'S VULNERABILITIES: The assessment also reveals critical weaknesses at the middle layers of the stack (Section VII).

Cloud infrastructure is Canada's most acute controllable vulnerability: extraterritorial legal reach means that Canadian data residency does not equal Canadian data sovereignty, and recent incidents have demonstrated that foreign providers can deny service entirely under geopolitical pressure. The compute hardware layer presents equally severe vulnerabilities that cannot be addressed domestically. Most advanced AI chips and components are designed and fabricated outside Canada, with extreme concentration among a handful of foreign suppliers. At the foundation model layer, a small number of American companies dominate enterprise AI usage, and Canada has only one domestic alternative in Cohere.

Strategic Options for Canadian AI Sovereignty:

1. Sovereign Cloud Infrastructure: Sovereign cloud infrastructure is the highest-priority strategic option presented in this paper. Cloud is Canada's most acute vulnerability, yet it is also the layer where realistic domestic solutions exist (Section VIII.B, Appendix B). Two primary models present distinct approaches to achieving cloud sovereignty:

»» Juridical cloud sovereignty (modelled on France and Germany) requires Canadian-owned and operated infrastructure that creates a legal air gap from foreign jurisdiction. This includes licensed operator arrangements where a Canadian entity operates foreign technology under Canadian control.

»» Contractual cloud sovereignty (modelled on Australia) achieves sovereignty through outcome-based controls — encryption keys held by Canadian entities, cleared personnel, and audit rights — that any provider can meet regardless of corporate nationality.

The appropriate model depends on the sensitivity of the workload and the institutional context governing it. One feasible set of minimum requirements could include self-hosted government infrastructure or juridical sovereignty for classified workloads; juridical sovereignty for sensitive personal and organizational data held by the government; and contractual sovereignty for private-sector data at the same level of sensitivity. Finally, for general business operations, which account for most cloud usage, market conditions should determine adoption.

For sovereign public cloud, pooling federal and provincial demand through arms-length Canadian sovereign compute providers, rather than fragmenting procurement across jurisdictions, would build critical mass for domestic providers and potentially narrow the sovereignty premium through economies of scale.

2. Additional AI Tech Stack Options: Strategic options extend beyond cloud to every layer of the AI tech stack (Section VIII):

»» **Compute hardware:** Canada could pursue a managed dependency strategy involving supply chain diversification across allied nations, bilateral assurance agreements linked to Canadian strengths in energy and critical minerals, multilateral semiconductor engagement, and contingency stockpiling for critical systems.

»» **Foundation model access:** Options could include establishing procurement preferences for Canadian providers, deploying open-source models on sovereign infrastructure as a strategic hedge, and diversification across model sources to avoid single-provider lock-in. Enterprise policies could address shadow AI, where most Canadian workers using AI rely on uncontrolled consumer tools rather than enterprise-grade alternatives.

»» **Data and data governance:** Options include introducing data governance frameworks that enable AI development while maintaining sovereignty and implementing measures to address Canadian content underrepresentation in global training datasets.

3. Cross-Cutting Enablers: Four policy mechanisms span multiple layers of the AI stack and require sustained attention to enable the strategic options outlined above (Section VIII.F). Procurement reform is needed to align government cloud and AI purchasing with sovereignty objectives, including updating the Government Cloud Framework with explicit sovereignty tiers and translating the Buy Canadian Policy into ICT-specific guidance. Workforce and security clearance requirements could be reassessed (including modernizing the security classification framework itself) to

accommodate increased demand for qualified staff to operate sovereign infrastructure. Data portability requirements should ensure that switching between cloud providers remains practically feasible, not just theoretically possible. And domestic sovereignty audit capabilities are needed to validate provider claims on an ongoing basis, ensuring compliance rather than one-time certification.

4. CUSMA Preparation: The scheduled CUSMA review in July 2026 presents both risk and opportunity for Canada's digital sovereignty position (Section VIII.G). The United States has signalled aggressive positioning against what it characterizes as digital trade barriers, and U.S. industry groups are advocating for constraints on Canadian sovereignty measures. Canada should defend national security exceptions and government procurement carve-outs as non-negotiable foundations for sovereign cloud and AI policy. Digital sovereignty provisions should not become bargaining chips for concessions in unrelated sectors.

5. Foundation Model Training and Domestic Research: Even with robust sovereign AI inference capacity, Canada remains vulnerable if every model it uses originates from foreign suppliers or if model makers revoke access or restrict functionality. We outline two key options (Section IX). The first would be to support Cohere (one of very few foundation model companies outside the United States and China) more explicitly as a national champion, ensuring it has the resources and government support to remain competitive as global competition escalates. The second would be for Canada to pursue multinational frontier AI partnerships with like-minded democracies, pooling compute and committing to open-source collaboration, because no single middle power can sustain frontier model training alone. Canadian researchers also face severe capacity constraints that risk driving talent abroad, in which case domestic research compute would also require expansion.

6. Strengthening State Capacity and AI Leadership: Building sovereign AI requires matching state capacity (Section X). Canada's digital and AI governance is dispersed across at least six institutional actors, and no single entity possesses both the strategic authority and the operational capacity to drive a coherent sovereign AI strategy. To treat sovereign AI as a strategic national priority, the federal government should consider consolidating authority into a single institutional vehicle, including, potentially, a fully resourced Ministry of Digital and AI with cross-government delivery authority and/or a dedicated Sovereign AI Unit with investment capital and a clear mandate to buy, build, and invest in major sovereign AI projects. Institutional reform should be complemented by treating digital government modernization as a core component of sovereign AI strategy, and by establishing federal-provincial-territorial coordination on AI and data governance.

Timelines: AI sovereignty will not be achieved overnight, but meaningful progress is achievable by 2030 with deliberate action. The strategic options outline both immediate steps — data sensitivity tiers, procurement reform, security clearance modernization, CUSMA preparation, confronting shadow AI policies, and AI state capacity — and longer-term investments that can be made in sovereign data centres, multinational partnerships, and research infrastructure. Sovereignty is a spectrum, not a binary. Each action that reduces Canada's exposure to foreign leverage strengthens the country's position.

Canada has the ingredients: world-leading researchers, abundant clean energy, a world-class foundation model company, a growing sovereign infrastructure ecosystem, AI firms, and democratic institutions worth protecting. The window to act is open, but opportunities will narrow as global investment decisions harden into long-term commitments and advantages for other nations.

I. INTRODUCTION

In January 2026, OpenAI's GPT-5.2 Pro solved Erdős Problem #728, a mathematical challenge that had stumped humans for decades, "more or less autonomously," in the words of Fields Medal-winning mathematician Terence Tao (Bastian, 2026). Six months earlier, both OpenAI and Google DeepMind achieved gold-medal standards at the International Mathematical Olympiad (Cai & Singh, 2025). AI systems now draft legal briefs, generate working code, diagnose medical images, and conduct scientific research, capabilities that were the stuff of science fiction a decade ago. This technology is now embedded in nearly every aspect of modern life, from the visible AI tools like ChatGPT, Claude, and Gemini that millions use daily, to the invisible algorithms that quietly shape our information, pricing, recommendations, and decisions.

The economic implications are profound: AI-related investment in the United States topped more than $100 billion in the first half of 2025, accounting for nearly 92% of U.S. GDP growth during that period (Oliver Wyman, 2026; Furman, 2025). The amount of private capital flowing into AI is comparable to similar industrial booms such as the build out of the railroads, electricity, and the internet. The infrastructure decisions being made today around where to build data centres, which platforms to adopt, and which standards to set will shape the technological landscape for a generation.

This technological transformation arrives amid a fracturing international order. Great power competition has made technology an increasingly central battleground. The U.S. now explicitly frames technological dominance in AI as essential to national security. Recent months have made abstract threats concrete for Canada: tariff threats, territorial provocations, and "51st state" rhetoric have clarified that the rules-based order we once relied upon has not just frayed, but ruptured.

Canada occupies a paradoxical position in this landscape. We are a classic middle power: a rules-follower, a multilateralist, an open economy. We are also the birthplace of modern AI. Geoffrey Hinton won the Nobel Prize in Physics, Yoshua Bengio and Richard Sutton continue to shape the field's direction, and Canadian institutions trained a generation of the world's leading researchers. Canadian researchers have contributed to the discoveries of neural networks, deep learning, and the transformers for large language models. Yet for all our contributions to AI's development, we control neither the major firms that dominate its deployment nor the key infrastructure that powers AI.

This dependency is a dangerous vector for coercion and exploitation. COVID-19 revealed what happens when critical capabilities reside elsewhere: early in the pandemic, Canada was forced to rely on other countries to produce vaccines at scale (Tasker, 2021). While we eventually responded with over CAN$2.5 billion in investments, and in September 2025, produced our first made-in-Canada mRNA doses (ISED, 2025a; ISED, 2025b; ISED, 2025c), it nevertheless revealed how vulnerable we were. Canada cannot afford to make the same mistake with AI.

The solution, to quote Canada's first Minister of Artificial Intelligence Evan Solomon, is to recognize that "sovereignty does not mean solitude" (Solomon, 2025). The goal is not digital isolationism, but rather strategic autonomy: ensuring Canadian data and infrastructure are governed by Canadian laws and values, secure from foreign threats, and resilient to geopolitical pressure.

This paper provides a pragmatic, risk-based framework for achieving that goal. No country can be fully self-sufficient in AI, not even the United States or China. The question for Canada is not whether to have dependencies, but how to structure them to preserve choice and reduce foreign leverage. Sovereignty carries costs: sovereign infrastructure may involve a price premium, and the most secure options may

Pullquote here if needed for visual interest and entry point, pullquote. Dere prerib volup solupiet parita simi

constrain access to cutting-edge capabilities. But failing to adopt AI is itself a sovereignty risk: a country that falls behind becomes economically weaker, less competitive, and strategically vulnerable. Our approach is to proceed layer by layer through the AI technology stack, distinguishing controllable vulnerabilities from uncontrollable ones, and identifying where intervention, particularly across government and regulated sectors where sensitivity is highest, can make a difference.

The paper begins by examining the strategic context that makes AI sovereignty urgent, which are the geopolitical and economic imperatives driving action. We then survey how other nations are responding, from American dominance strategies to Chinese self-sufficiency to European regulatory approaches and middle power alternatives. A technical foundation follows, explaining the architecture of modern AI and the critical distinction between training and inference.

With this groundwork established, we define digital sovereignty, extend the idea to AI sovereignty, assess Canada's position across each layer of the technology stack, and present strategic options spanning the full stack: from procurement reform and data governance to sovereign cloud infrastructure and foundation model strategy.

The options we present are a menu: different interventions can be pursued independently, at different speeds, and in different combinations. In practice, AI sovereignty is a spectrum, not a binary, and each step that reduces Canada's exposure to foreign leverage strengthens the country's position. Some measures can have an immediate benefit, while others will require sustained effort to pay dividends. While it will take time, we believe meaningful progress is achievable by 2030.

Canada has the ingredients to thrive in this increasingly AI-driven world: world-leading researchers, abundant clean energy, a foundation model company, a growing sovereign infrastructure ecosystem, and democratic institutions worth protecting. The window to act is open but closing as investment decisions harden into long-term commitments. The question is whether we will make the necessary choices, and make them in time.

II. THE STRATEGIC CONTEXT: WHY AI SOVEREIGNTY MATTERS NOW

II.A The Geopolitical Imperative

Technology has become an increasingly dominant factor in geopolitical rivalry and great power competition. This is not new, as industrial capacity and technological advantage have shaped the balance of power for centuries, but the emergence of artificial intelligence represents something qualitatively different. AI has the potential to impact national security across every domain: cybersecurity capabilities, cutting-edge research and development, economic productivity, and the ability to design the most sophisticated defence systems. As AI becomes increasingly powerful, control over the technology stack that enables it becomes inseparable from national power itself.

Moreover, the prospect of AI systems matching or exceeding human cognitive abilities across a wide range of domains—what some call "Artificial General Intelligence"—could further shift geopolitical power dynamics. A nation that achieves a significant lead in advanced AI capability would be well-positioned to compound that advantage, using superior AI to accelerate subsequent advances and widen the gap with competitors.

The Trump administration's "America First" National Security Strategy, released in November 2025, marks an inflection point in global affairs. Unlike previous strategies that balanced American interests with alliance commitments, this document places economic power, industrial capacity, and sovereignty at the centre of U.S.

national security policy. This “America First” doctrine emerges against the backdrop of a crumbling international system, where the multilateral world order that structured post-Cold War relations has given way to naked great power competition. The Russian invasion of Ukraine in 2022 was the most visible rupture, as a permanent member of the UN Security Council launched a war of territorial conquest in Europe, but it was not the last. The willingness of major powers to intervene anywhere, override international institutions, and impose their will through force has become normalized.

Technology dominance, specifically in artificial intelligence, biotechnology, and quantum computing, sits at the heart of U.S. strategy. The National Security Strategy explicitly frames this as essential to both economic security and homeland protection: American technology and American standards must "drive the world forward" (The White House, 2025b). This is not aspirational rhetoric but operational doctrine, with technology sales and defence purchases identified as transactional foreign policy tools to tip global influence in America's favour.

Within this framework, U.S.-based technology giants have become central instruments of American geopolitical power. Their platforms are not neutral infrastructure but carriers of American norms, American politics, and American jurisdiction. As their technologies become global standards, they bring American law with them, allowing the United States to hardcode its technical choices into geopolitical leverage. Big Tech is not just an economic actor; it is increasingly becoming a central vector of American power.

China has emerged as the other AI superpower, locked in strategic competition with the United States across every dimension. China is investing heavily in a domestic AI ecosystem precisely to avoid dependence on American technology, and, by extension, American power and norms. And China’s efforts have been paying off: the performance gap between American and Chinese AI models narrowed from 9.3% to 1.7% in a single year (JPMorgan Chase, 2025). This is not a gradual evolution, but a rapidly changing environment driven by fierce competition.

The concentration of AI capacity is stark. The United States controls approximately 75% of global AI compute capacity; China controls 15%; the entire G7 minus the United States controls just 6% (Abecassis et al., 2025). Chip supply has dominated the narrative around China and Taiwan for precisely this reason: control over semiconductor production is control over AI's physical layer. This concentration, alongside concentrations of talent, data, and model ownership, means the strategic landscape is hardening rapidly.

Control over the AI technology stack has become inseparable from sovereignty itself. Whether chips, foundation models, compute infrastructure, or the human capital that develops these systems, dependency on any critical layer creates leverage for foreign powers.

II.B The Economic Imperative

AI has the potential to be the most significant economic transformation since the computing revolution, with capital deployment at historically unprecedented levels and credible projections of productivity gains that could reverse decades of stagnation. Several AI-driven factors are now reshaping the global economy in ways that no advanced nation can afford to ignore: massive levels of private investment, dramatic shifts in corporate valuations, substantial productivity potential, the emergence of entirely new economic categories, and uncertain but consequential impacts on employment.

The scale of current AI investment is extraordinary and accelerating. Hyperscaler capital expenditure from Alphabet, Microsoft, Amazon, and Meta totaled approximately $237 billion in 2024 and is projected to reach $540 billion in 2026 (Goldman Sachs, 2026). McKinsey projects $3.7-7.9 trillion in global AI infrastructure investment through 2030 (Noffsinger et al., 2025). Total AI-related investment accounted for nearly 92% of U.S. GDP growth in the first half of 2025. Without tech infrastructure spending, annualized GDP growth would have been just 0.1% (Furman, 2025). AI investment has become the primary engine of economic expansion in the world's largest economy.

The equity market impacts have been equally dramatic. NVIDIA, which controls over 90% of the AI hardware market, has surpassed $4.5 trillion in

market capitalization, making it one of the most valuable companies in history (Lewsing, 2025a; Lewsing, 2025b). The gap between AI leaders and laggards widened by 60% between 2020 and 2023 (Hall et al., 2024). These valuations reflect market expectations of transformative productivity gains across the economy. Companies spent $37 billion on generative AI in 2025, up from $11.5 billion in 2024 — a 3.2 times increase in a single year (Menlo Ventures, 2025).

The projections of potential impact on productivity, though spanning a wide range, are substantial. Goldman Sachs projects generative AI could raise global GDP by 7%, which is approximately $7 trillion, and boost U.S. labour productivity by 1.5 percentage points annually over a decade (Goldman Sachs, 2023). More conservative academic estimates from Daron Acemoglu suggest approximately 0.07 percentage points of annual total factor productivity growth (IMF, 2025). The OECD provides the most methodologically rigorous country-specific analysis: for Canada, projections show 0.35 percentage points in annual productivity growth under slow adoption and 1.13 percentage points under rapid adoption (OECD, 2025). The gap between slow and rapid adoption represents a cumulative difference of roughly 8% of GDP over a decade, or approximately CAN$250 billion to $300 billion in higher GDP (authors' calculations).

Beyond productivity gains in existing enterprises, AI is creating entirely new categories of economic activity. Global venture capital investment in AI exceeded $100 billion in 2024, an 80% increase from the prior year, with nearly one-third of all global venture funding directed to AI companies (Teare, 2025). Among 54 new unicorns in 2025, more than half are AI companies (Teng & He, 2025). New categories are emerging rapidly: vertical AI became a $3.5 billion investment category in 2025, triple the prior year; AI coding tools attracted $4 billion in investment, up from $550 million; and AI agent startups raised $3.8 billion in 2024 (Menlo Ventures, 2025; CB Insights, 2025).

The employment effects remain uncertain but will likely be consequential. Short-term data shows net job creation in the United States, with approximately 119,900 direct jobs created in 2024 versus 12,700 lost, but the positive ratio is heavily weighted by data centre construction (Ostertag, 2025). More concerning are emerging signs of displacement concentrated among young workers: Stanford researchers found employment for workers aged 22-25 in AI-exposed occupations declined 13% relative to less exposed roles since 2022 (Brynjolfsson et al., 2025). The Federal Reserve Bank of St. Louis warns that "we may be witnessing the early stages of AI-driven job displacement." The notable difference with AI is that it is affecting "white collar" jobs in industries like management consulting, banking, and public administration (Ozkan & Sullivan, 2025). In 2024, Statistics Canada conducted widely cited research on "Potential Artificial Intelligence Occupational Exposure in Canada," which estimates the job impacts across a variety of sectors (Statistics Canada, 2024).

The scale and pace of this economic transformation, from investment levels to productivity potential to new industry creation, represents a structural shift that will reshape competitive advantage among nations for decades to come.

II.C Why Focus on AI Sovereignty Now

The geopolitical stakes and economic opportunity of artificial intelligence are clear. But why must Canada act now? Three factors make the present moment decisive: the rapid pace of technological change, the risk of lock-in as billions of dollars of new investment are deployed over the next half decade, and a closing window of opportunity to influence the direction of AI globally.

AI capabilities are advancing at an extraordinary pace, and the advantages compound rapidly for early movers. Nations and firms that establish strong positions in compute infrastructure, model development, and talent pipelines will use those advantages to accelerate subsequent progress. Those that fall behind will find the gap widening rather than narrowing (Abecassis et al., 2025). Thus, the choice to defer action is itself a choice; a choice to accept dependence on foreign powers whose interests may not align with Canada's.

The scale of current investment amplifies this dynamic. As mentioned above, hundreds of

billions of dollars are being deployed annually to build data centres, train foundation models, and establish the infrastructure that will underpin AI-driven economies for decades (Goldman Sachs, 2025). The decisions being made today around where to locate infrastructure, which platforms to adopt, and what technical standards to embrace will shape the technology landscape for a generation. Choosing platforms without regard for sovereignty implications risks locking Canada into dependency that will be difficult and costly to reverse.

Yet this very dynamism also creates opportunity. The architecture of tomorrow's AI systems is not yet settled. Investment is still being allocated. Technical standards are still being written. A deliberate, sovereignty-oriented approach can shape how these resources are deployed, capturing the economic benefits of AI adoption while preserving the capacity for autonomous action. The question is not whether Canada will participate in the AI transformation, but whether it will do so on its own terms. The policies advanced today will determine the answer.

III. THE GLOBAL LANDSCAPE: INTERNATIONAL APPROACHES TO AI SOVEREIGNTY

The geopolitical rivalry and economic transformation spurred by AI has resulted in nations pursuing distinct, and increasingly urgent, approaches to AI sovereignty, shaped by their strategic positions, industrial capacities, and political systems. The United States seeks to project its technological dominance globally. China pursues comprehensive self-sufficiency across the AI stack. The European Union attempts to balance regulatory leadership with industrial capacity-building. For middle powers, the choices made by these three blocs define the strategic terrain they must navigate. Understanding these approaches is essential before Canada can chart its own course.

III.A United States: Projecting AI Dominance

Vice President JD Vance's keynote address at the Paris AI Action Summit in February 2025 crystallized American AI policy. He asserted that America possesses "all components across the full AI stack, including advanced semiconductor design [and] frontier algorithms," positioning the United States as the indispensable provider of AI technology to the world (Vance, 2025). The White House's AI Action Plan, released later in July 2025, makes explicit what Vance's speech implied: the objective is "unquestioned and unchallenged global technological dominance" in artificial intelligence (The White House, 2025a). The plan explicitly frames AI development as a race against strategic competitors, particularly China, with the central premise that "whoever has the largest AI ecosystem will set global AI standards and reap broad economic and military benefits" (The White House, 2025a).

Despite very stark stylistic differences, both recent U.S. administrations have pursued strategies of American AI dominance. The Trump approach is overt: the November 2025 National Security Strategy frames AI as a "core, vital national interest," with American technology and standards to "drive the world forward" (The White House, 2025b). The July 2025 "America First" AI Action Plan pivoted from applying restrictions to export promotion. The U.S. goal now is to export complete AI ecosystems to create dependency rather than simply denying technology to competitors (Wise et al., 2025; Cook et al., 2025).

The Biden approach was less visible but equally serious. Then-National Security Advisor Jake Sullivan declared in September 2022 that securing "as large a lead as possible" in AI was a national security imperative (Harithas & Schumacher, 2024). The resulting export controls operated in cascading sequence: block direct access to advanced semiconductors, then

deny access to tools needed to fabricate chips domestically, then obstruct access to critical components required to build manufacturing machinery. The sequence was designed, in the words of one analysis, to "trap China in a technological cul-de-sac and sustain U.S. dominance" (Harithas & Schumacher, 2024).

American technology giants serve as vectors for this dominance strategy. Their platforms are not neutral infrastructure but carriers of American norms and American jurisdiction. American law compounds this leverage: the CLOUD Act enables U.S. authorities to compel disclosure of data from American companies regardless of where data is stored, and FISA Section 702 allows surveillance of foreign users on American platforms without their knowledge (Avasant, 2025; Ortiz, 2025). A preview of coercive leverage became evident when a French judge serving on the International Criminal Court found himself digitally cut off from American platforms after U.S. sanctions targeted the court, demonstrating that even close allies are not immune (Kirchner, 2025).

This combination of private dominance with state leverage creates what Turobov (2025) terms a "kill switch" vulnerability. Export controls, cloud service terms, and sanctions frameworks explicitly reserve Washington's right to terminate access to critical technologies through executive action, with no recourse for affected governments.

Commerce Secretary Howard Lutnick has explicitly demanded the European Union balance their tech rules (likely referring to the AI Act, Digital Services Act, and Digital Markets Act) in exchange for lower steel tariffs, framing European regulation as protectionist non-tariff barriers (Valero et al., 2025). The very regulatory frameworks other nations develop to protect their citizens now face direct American contestation.

III.B China: Building a Self-Sufficient AI Stack

China pursues comprehensive state-directed sovereignty under Xi Jinping's directive for an "independent and controllable" AI ecosystem (Pomfret & Zhen, 2025). This approach is consistent with China's broader industrial strategy to achieve self-sufficiency in emerging technological and industrial domains. Although not yet fully realized, China may be the only country with the medium-term capacity to build a completely domestic AI stack, from raw materials and rare earths refinement through to semiconductors, data centres, and foundation models.

The January 2025 release of DeepSeek-R1 marked a watershed moment. Initially, reports stated that DeepSeek's model was built for under US$6 million (Soni & Kachwala, 2025), a fraction of typical training costs. However, those figures have been heavily contested, with some analysts suggesting the realistic training and associated capital expenditures were more likely US$1.6 billion (Patel et al., 2025). Nonetheless, given the fraction of global compute spend that either of these figures represent, sovereign AI capability is more achievable than previously assumed (KEYSS Inc, 2025; Jiao, 2025). DeepSeek's success prompted one of the largest single-day market value losses in history for American technology stocks, with NVIDIA alone losing nearly $600 billion in market capitalization (Subin, 2025), signalling that American AI dominance was not guaranteed. Some have even questioned whether China is quietly winning the AI race (Jamali, 2026).

China's semiconductor push represents its most capital-intensive sovereignty investment. Big Fund III dedicated $47.5 billion, exceeding the U.S. CHIPS Act's $39 billion. Shanghai targets 70% of data centre chips domestically designed or produced by 2027; Beijing targets complete independence (TrendForce, 2025a). Despite extreme ultraviolet (EUV) lithography restrictions, SMIC has produced 7nm chips through advanced DUV multi-patterning techniques (TrendForce, 2025b), and Chinese scientists completed a working EUV lithography prototype in early 2025 by reverse-engineering ASML's technology (TrendForce, 2025c).

China's adoption of AI is outpacing America's: 83% of Chinese decision-makers report using AI compared to 65% in the United States, reflecting state-directed deployment across the economy (JPMorgan Chase, 2025). Chinese large language models' share of global usage surged from 1.2% in late 2024 to nearly 30% in 2025 (Xu, 2025). Alibaba's Qwen family reached 700 million downloads on Hugging Face, surpassing Meta's

Llama, with over 170,000 derivative models created by developers worldwide (Jiang, 2025a). The pricing advantage is substantial, with Chinese models often being cheaper and open source. (Hill, 2026). Even Airbnb CEO Brian Chesky chose Qwen over ChatGPT because it is "fast and cheap" (Jiang, 2025b).

This open-source strategy serves geopolitical as well as commercial purposes. Researcher Qiang Xiao, speaking at the Montreal International Security Summit in October 2025, described the phenomenon as "infrastructure colonization." The broad adoption of Chinese large language models embeds foreign political assumptions into the architectures of software, workflows, and public knowledge systems (Lamensch, 2026). Research comparing ChatGPT and DeepSeek finds systematic differences: American models anchor explanations in international law, multilateral institutions, and democratic norms; Chinese models emphasize state sovereignty, national unity, and geopolitical stability. Uganda launched "Sunflower," a large language model (LLM) built on Qwen architecture; Malaysia introduced NurAI, the world's first Sharia-aligned LLM, refining Chinese foundations. In Brazil, India, and Nigeria, lightweight downloadable Chinese models are becoming the default choice. These open-source offerings represent an alternative to lock-in with American providers, while potentially embedding Chinese values and norms instead.

Pullquote here if needed for visual interest and entry point, pullquote. Dere prerib volup solupiet parita simi

China's data sovereignty framework complements its industrial policy. The Personal Information Protection Law, Network Security Law of the People's Republic of China, and Data Security Law create comprehensive data localization requirements and explicit extraterritoriality provisions (Yun, 2025; Xie et al., 2024). Data produced in China is treated as a state asset under exclusive jurisdiction, with foreign firms facing joint venture requirements and forced technology transfer as conditions of market access.

III.C Europe: Regulatory Leadership, Uncertain Results

The European Union pursues a distinctive "third way," combining regulatory leadership with industrial capacity-building. However, the gap between ambition and capability remains stark, and recent developments suggest growing uncertainty about the strategy's viability. Europe currently imports over 80% of digital technologies; U.S. firms control 70% of European cloud infrastructure, and 70% of foundational AI models come from the U.S. (Bria et al., 2025).

The regulatory apparatus is substantial. The EU AI Act, which entered force in August 2024, represents the world's first comprehensive legal framework for AI. It works "hand-in-glove" with the General Data Protection Regulation to establish trustworthiness requirements across risk categories (Clark et al., 2024). Yet the regulatory strategy faces both internal retreat and external pressure. The General-Purpose AI Code of Practice (European Commission, 2026), published in July 2025, was created to simplify obligations for major AI firms. Despite this effective carve-out, the November 2025 Digital Omnibus proposal delays key AI Act obligations for high-risk AI systems by up to 16 months, to December 2027. This is clearly a response to former Italian Prime Minister Mario Draghi's competitiveness report warning that Europe risks missing the paradigm shift (Bracy, 2025; Hickman et al., 2025). The delay followed intense industry lobbying: more than 40 European CEOs demanded a two-year "clock-stop" on the AI Act in July 2025 (Kroet, 2025).

Infrastructure initiatives aim to close the capability gap. The proposed EuroStack represents a commitment of 300 billion euros over 10 years for federated digital infrastructure spanning seven technology layers, from raw materials and chips through cloud infrastructure to data and AI applications (Bria et al., 2025). Ahead of the February 2025 AI Action Summit, France announced they would invest 109 billion euros into AI infrastructure (Browne, 2025). Yet dependencies persist: GAIA-X, launched to ensure European digital sovereignty, faced criticism for incorporating the very American cloud providers it ostensibly sought to contest, prompting accusations it became "a Trojan horse for Big Tech in Europe" (Fermigier & Franck, 2020).

Mistral AI, a French company, provides concrete evidence that European AI capacity is achievable, though scale gaps persist. Founded by alumni of DeepMind and Meta, Mistral reached a valuation of approximately $14 billion in September 2025, having raised over 3 billion euros across seven funding rounds in just 29 months (Loizos, 2025). President Emmanuel Macron personally endorsed the company in February 2025, urging French citizens to "download Le Chat, which is made by Mistral, rather than ChatGPT." The app reached one million downloads in two weeks (Dillet, 2025). Mistral's open-source approach and European data sovereignty positioning have made it dominant in European enterprise deployments where regulatory compliance outweighs raw model performance. Yet Mistral remains an exception rather than a pattern: European AI companies secured 55% more investment in Q1 2025, but the absolute sums remain far below American and Chinese levels (Loizos, 2025).

III.D Middle Powers: Navigating Between the Dragon and the Eagle

Middle powers face a common structural challenge in the AI era. The most capable cloud platforms, foundation models, and AI accelerator chips are overwhelmingly controlled by either American or Chinese firms, yet full technological self-sufficiency is beyond any middle power's reach. Middle powers are thus forced to develop strategies that balance dependencies on foreign sources with the imperative to avoid falling behind technologically and economically.

A comprehensive survey of middle power AI strategies is beyond the scope of this paper. The approaches of nations like France, Germany, the United Kingdom, Japan, South Korea, Australia, and the Gulf states vary enormously in ambition, mechanism, and political context. What follows instead highlights a few notable developments that are instructive for Canada.

Cloud sovereignty is perhaps the most consequential policy innovation among middle powers: specifically, in how governments manage the risk that sensitive national data processed on foreign-owned platforms may be subject to foreign legal orders. Two instructive models have emerged.

France, for example, has pioneered a model that creates a legal air gap between foreign technology and foreign jurisdiction. Under France's Cloud de Confiance doctrine, sensitive government data may only be processed on cloud services qualified by ANSSI, the national cybersecurity agency, under a standard whose defining feature is insulation from non-EU extraterritorial laws (Capgemini, 2024). Germany has adopted similar frameworks, resulting in the creation of domestic entities like Delos Cloud to meet sovereign cloud requirements (European Cloud, 2025).

Australia has taken a different path, defining sovereignty not by provider nationality but by contractual and operational controls. Its Hosting Certification Framework is origin-neutral: any provider, including American hyperscalers, can serve sensitive government workloads provided it accepts binding requirements around data residency, personnel clearances, supply-chain transparency, and government oversight rights. AWS, Google Cloud, and Oracle have all achieved "Certified Strategic" status (the framework's highest tier) demonstrating that foreign ownership is not a barrier for handling Australia's sensitive unclassified workloads. However, the framework is not static: in November 2025, the Department of Home Affairs paused new certifications as part of a broader reform process, signalling that even Australia's comparatively permissive model is tightening as sovereign-cloud expectations evolve (Horton, 2024; Holding Redlich, 2021; Oracle, 2024; Department of Home Affairs, 2025).

Infrastructure investment, including in both cloud infrastructure and semiconductor supply chains, represents the most common response. The U.K.'s July 2025 Compute Roadmap commits 2 billion pounds, including over 1 billion pounds to expand the AI Research Resource twentyfold by 2030 (Baker Botts, 2025). Japan's 1 trillion yen (US$6.34 billion) Sovereign AI package addresses decades of "digital deficit." Japan's deficit includes heavy reliance on foreign software and cloud infrastructure (SCMP, 2025). South Korea's $534 billion semiconductor strategy seeks to transform its memory chip-focused industry into a comprehensive innovation ecosystem across the full value chain (Hae-rin, 2025).

Regulatory pragmatism distinguishes middle power approaches from the EU's comprehensive framework. Japan's May 2025 AI Promotion Act implements an "innovation-first" blueprint that deliberately avoids stringent rules, with a focus instead on cooperation and reasonable efforts from companies, as well as "name and shame" enforcement (Paulger, 2025). Australia's December 2025 National AI Plan applies existing technology-neutral laws rather than creating new AI-specific regimes (similar to the U.K.). This is a bet on regulatory agility over more comprehensive frameworks that may constrain adoption (Bird & Bird, 2025; Boyle et al., 2025).

For Canada, these international approaches illuminate both constraints and opportunities. The country cannot match U.S. or Chinese scale, but also cannot afford to imitate Europe's widening productivity gap. What Canada can do is craft an approach for strategic autonomy that manages dependencies while preserving the capacity to act independently when national interests require it. We can also partner with countries, particularly other middle powers, to advance our chosen strategic options.

IV. THE ARCHITECTURE OF MODERN GENERATIVE AI

Before assessing Canada's strategic options, we must understand the underlying technology. This section provides a primer on generative AI: what these systems can do today, how developers build them, and who controls their development. The distinctions introduced here, particularly between training and inference and between frontier, small, and open-weight models, are essential for understanding which sovereignty interventions Canada can feasibly pursue.

IV.A Current Capabilities and the Pace of Advancement

Today's AI systems perform a wide range of knowledge work tasks, including drafting documents, writing code, analyzing data, and answering complex questions at levels that approach or exceed human performance in many contexts. AI agents complete software engineering tasks that would take humans hours to accomplish. Frontier models now score at or above expert-human levels on widely used knowledge benchmarks spanning dozens of professional and academic domains, including law, medicine, and the sciences (Maslej et al., 2025).

These capabilities improve rapidly and unpredictably. According to Sastry et al. (2024), the compute used for training notable AI systems has doubled approximately every six months since 2010. This is an approximately 350 million-fold increase over 13 years that dramatically exceeds the roughly two-year doubling time of Moore's Law. Moreover, traditional benchmarks designed to challenge AI systems are being surpassed within months of introduction. The AI Index 2025 Report by Stanford University reports that scores rose by 18.8, 48.9, and 67.3 percentage points on image-and-text reasoning (MMMU), graduate-level science (GPQA), and real-world software engineering (SWE-bench) benchmarks respectively within a single year of their introduction (Maslej et al., 2025).

Artificial General Intelligence

Discussions of AI increasingly reference "artificial general intelligence" (AGI): hypothetical systems capable of matching or exceeding human cognitive abilities across all domains. Many experts project AGI-level systems within five to 10 years, though such claims remain highly uncertain (Abecassis, 2024). This report does not attempt to predict AGI timelines or assess speculative long-term scenarios. The analysis that follows focuses on current capabilities and near-term developments; the systems being deployed today and the infrastructure decisions that will shape Canada's position over the next decade. Whether or not AGI emerges in the near term, the strategic choices Canada faces regarding existing AI technology remain pressing and consequential.

The METR benchmark measures the length of tasks AI agents can autonomously complete. The "50% time horizon" (tasks at which models succeed half the time) has doubled approximately every seven months from 2019–25, though the rate accelerated to four months during 2024–25. Current frontier models like GPT-5 achieve a time horizon of approximately two hours and 17 minutes. If recent trends continue, AI systems could autonomously complete 16-hour tasks by February 2027 and five-day tasks by April 2028 (METR, 2025).

Declining costs provide a crucial complement to rising capabilities: the cost of achieving GPT-3.5-equivalent performance dropped 280-fold between 2022 and 2024 (Maslej et al., 2025; Noffsinger et al., 2025). These reductions stem from improvements across multiple dimensions: more efficient model architectures, better hardware utilization, optimized inference techniques, and economies of scale in deployment. However, developers have historically consumed these efficiency gains by scaling to larger models rather than reducing aggregate spending. This is the Jevons paradox in the context of AI compute (Brookfield, 2025). Any AI strategy premised on current technical limitations must account for rapid, potentially discontinuous capability growth.

IV.B Training vs. Inference: A Critical Distinction

Within AI systems, the distinction between training and inference is crucial for sovereignty analysis. Training is the computationally intensive process of creating a model by processing vast datasets; a massive, one-time capital expenditure where a single frontier training run uses thousands of specialized chips for months. Inference is running a trained model to generate outputs for users, which is an ongoing operational cost that scales with usage.

Inference is what powers every AI application in use today. When a user interacts with ChatGPT, queries a customer service chatbot, or receives coding suggestions from GitHub Copilot, they are consuming inference. Microsoft's Copilot integration across Office products, AI-powered search engines, content moderation systems, and enterprise document analysis tools all run on inference. Every prompt, query, and API call consumes compute at the inference layer. (Note: Fine-tuning, which is adapting an existing model to specific domains or tasks at a fraction of pre-training cost, sits in between training and interference.)

Table IV.2: Training vs. Inference Characteristics

Dimension	Training	Inference
Cost profile	Massive upfront CAPEX ($100 million –$1 billion+ for frontier and growing)	Ongoing OPEX (scales with usage)
Who does it	Handful of frontier labs globally	Any organization using AI
Geographic concentration	Highly concentrated (mostly U.S. & China)	Distributed; benefits from user proximity
Project share of frontier AI model lifetime cost	10–20%	80–90%
Data sovereignty exposure	Training data (often web-scraped)	User data flows through inference

Crucially, the economic significance shifts decisively toward inference as AI deployment scales. AWS estimates 90% of its AI workload is inference (Patterson et al., 2021) and Brookfield estimated roughly 75% of future AI compute demand will come from inference by 2030 (Brookfield, 2025). Google reports needing to double AI serving capacity every six months to meet demand, with projections of 1,000-times growth in four to five years (Elias, 2025). Agentic AI, which chains dozens of model calls to accomplish a single goal, will multiply inference needs even further. Thus, despite escalating training costs, it is expected that the majority of investment in AI compute over the next decade will be for inference.

IV.C Frontier Models and the Cost of Development

A foundation model is a large AI system trained on broad data that can be adapted to a wide range of downstream tasks. These models, sometimes called "base models" or large language models, represent the core technological asset in modern AI: the product of massive investment in compute, data, and research that other developers then build upon. Foundation models matter because they determine what AI systems can do. An application built on a capable foundation model inherits that capability; one built on a weaker model faces fundamental limitations that no amount of fine-tuning can overcome.

Foundation models are difficult to replicate because they build upon decades of accumulated research breakthroughs, particularly in deep learning and transformer architectures, combined with the engineering knowledge to train systems at scale. The organizations capable of developing frontier foundation models are listed in Table 1. This small group reflects not just enormous capital requirements but the concentration of specialized expertise required for frontier model development.

Table IV.1: The Frontier Model Landscape (Early 2026)

Developer	Headquarters	Current Frontier Model(s)	Open Source?
OpenAI	United States	GPT-5.2	No
Anthropic	United States	Claude 4.6 (Opus, Sonnet)	No
Google DeepMind	United States/ United Kingdom	Gemini 3.0	No
Meta	United States	Llama 4 series	Yes (Open weight)
xAI	United States	Grok 4.1	No
DeepSeek	China	DeepSeek-V3, R1	Yes (Open weight)
Alibaba	China	Qwen3	Yes (Open weight)
Mistral	France	Mistral Large 3	Yes (Open weight)
Cohere	Canada	Command A	No

Competition to remain at the technological frontier is fierce. "Scaling laws" — empirical relationships showing that model performance improves predictably with more compute, data, and parameters (Sastry et al., 2024) — have driven exponential increases in training costs. Costs per training run had escalated to $500 million by mid-2024, with projections suggesting costs will exceed $2 billion by 2026 and potentially reach $6.6 billion per training run by 2028 (Noffsinger et al., 2025; Abecassis et al., 2025; Pilz et al., 2025; Cottier et al., 2024; Emberson & Somala, 2025). Competition for elite researchers and capital is intense. By 2027, frontier AI model training may be too expensive for all but the most well-funded organizations, with the largest training runs potentially costing more than $1 billion (Cottier et al., 2024).

Canada maintains a presence in this landscape through Cohere, though not at the same frontier scale as the largest U.S. labs. The concentration of frontier AI development creates what Abecassis et al. (2025) describe as "natural monopoly tendencies and widening moats. The extreme fixed costs combined with low marginal serving costs create winner-take-most markets." However, there are emerging signs that small language models can be more effective in specialized domains. Further, powerful, low cost and widely accessible open-weight models (particularly from Chinese developers) are continuing to invite questions about how strong the moats of frontier AI model companies like OpenAI and Anthropic are.

The Small Model Alternative

It's critical to note that not every application requires frontier-scale capability. A parallel ecosystem of small language models (SLMs) offers a different trade-off: narrower but sufficient capability for specific tasks, deployable on modest infrastructure. These models run on a single GPU or even consumer hardware, with lower inference costs critical for agentic workflows involving dozens of model calls per task. They suit well-defined tasks like document processing, summarization, classification, translation, and code assistance, and developers can fine-tune them for domain-specific performance that rivals larger models on narrow applications.

Many small open-source models are produced by the same companies that develop large foundation models: Meta, Mistral, Alibaba, Anthropic, and others release both frontier and small variants. These smaller models are often "distilled" from larger foundation models, transferring capability through techniques that compress the knowledge of a large model into a more efficient form.

The small language model market is projected to grow from $0.93 billion in 2025 to $5.45 billion by 2032 (MarketsandMarkets, 2025). In specialized domains, such as healthcare, SLMs can sometimes have higher accuracy than large language models. They can also be much cheaper (Invisible Technologies, 2025). SLMs can, in some cases, make sovereign AI efforts more economically and technically feasible. And increasingly in agentic AI applications, where a single user task may trigger many inference calls, reducing latency and per-query cost are critical factors.

IV.D Open-Source and Open-Weight Models

A significant ecosystem of openly available models exists, spanning capabilities from small efficient models to near-frontier performance. Understanding this landscape matters because open-weight models offer different sovereignty trade-offs than proprietary alternatives.

The terminology is important: "open weights" (model parameters available for download) differ from fully "open source" (weights plus training data plus code plus methodology). Vincent et al. (2025) identify a spectrum of openness, summarized in Table 3.

Table IV.3: The Spectrum of Model Openness

Level	What Is Released	Transparency	Examples
Fully Open	Weights + training code + documentation + pre-training data	Complete reproducibility; full scrutiny possible	OLMo (Ai2), Salamandra (BSC)
Open Weights with Closed Data	Weights + training software; pre-training data not released	Partial transparency; cannot fully audit training	Mistral models
Open Weights Only	Weights with limited transparency	Minimal transparency; cannot verify training process	Meta's Llama

The performance gap with proprietary models has narrowed and varies with virtually every new frontier model released: the Stanford AI Index 2025 Report reports the gap narrowed from 8.04% to just 1.70% on Chatbot Arena benchmarks in a single year (Reuel, 2025). The U.K. AI Security Institute estimates the capability lag between proprietary and open-weight models at only four to eight months (UK AISI, 2025; Artificial Analysis, 2026; Kwa et al., 2025). Current open-weight leaders include DeepSeek-R1 (DeepSeek, 2025), Qwen3-235B (Yang et al., 2025), and the Llama 4 series (Meta AI, 2025). DeepSeek achieved GPT-4-comparable performance at approximately 75% lower inference cost through efficiency innovations (Brookfield, 2025).

Open-weight models could offer three advantages for AI sovereignty efforts: deployment flexibility (potential to run on sovereign infrastructure without API dependencies), customization (potential to fine-tune for Canadian contexts, including French or Indigenous languages), and reduced vendor lock-in (providing optionality if proprietary providers become unavailable).

However, open models also have significant limitations. Open-weight models still typically originate from foreign countries, primarily the United States, France, and China. Thus, "open" does not eliminate foreign dependency. As mentioned above, open models can often lag the capabilities of closed, proprietary frontier models and this gap could "grow due to national security restrictions" (Abecassis et al., 2025). Licensing restrictions also sometimes apply: for example, Meta's Llama initially prohibited most military applications. Further, as the strategic AI landscape changes, companies may shift away from open-source commitments, limiting the supply of new, frontier-adjacent models. Meta has recently signalled a potential shift away from releasing open-weight models (Bellan, 2025). As Nguyen (2025) notes, an open-weight model is "epistemically plastic" in that it can be retrained, audited, or forked; but Abecassis et al. (2025) caution that "defence planning and critical infrastructure should not rely on systems where legal access may be revoked or denied in the first place."

"Public AI"

Beyond open-source models, a growing movement argues that genuine AI sovereignty requires what Marda et al. (2024) term "Public AI": a robust ecosystem of initiatives that promote public goods, public orientation, and public use throughout AI development and deployment. Under this framework, open weights are a necessary but insufficient condition for sovereignty. What matters equally is whether models are developed with public accountability, trained on transparent and appropriately governed data, and deployed on publicly accessible infrastructure.

Marda et al. (2024) estimate that over US$850 million has been invested in public AI labs over the past five years, including the Allen Institute for AI (Ai2) in the United States, Kyutai in France, and AI Singapore, though this remains a fraction of private sector investment. Critically, the Public AI concept addresses a gap in the open-weight model landscape: while commercial firms like Meta release model weights, market incentives mean they will only pursue the commercially profitable slice of AI research and development, neglecting smaller, customizable models for lower-resourced languages and contexts (Marda et al., 2024).

For Canada, this distinction matters because a sovereignty strategy built solely on using foreign open-weight models remains structurally dependent on the strategic choices of foreign corporations. A Public AI approach would complement open-weight adoption with domestically governed compute, data commons, and purpose-built models for Canadian public interest applications. We endorse the principles underlying Public AI and believe they should be pursued wherever feasible, but the policy recommendations in this report are calibrated to the economic and national security realities that Canada faces today; realities in which pragmatic engagement with commercial and open-weight ecosystems remains essential.

The distinctions established in this section, between training and inference, frontier and accessible capability, and open and closed models, form the foundation for understanding which sovereignty interventions are feasible. Canada cannot realistically compete across all dimensions of frontier model training. But Canada could build sovereign inference capacity, leverage Canadian-built and other open-weight models strategically, and participate in collective approaches to training where the economics of going alone are prohibitive.

V. DEFINING DIGITAL SOVEREIGNTY

AI intersects with fundamental questions of political authority and national control. This section establishes the conceptual foundation for the remainder of the report: what digital sovereignty means, how it can be compromised, and the challenges that middle powers face in navigating between dependency and isolation. The final subsection then applies this framework to the specific context of AI sovereignty.

V.A Sovereignty and Its Digital Extension

Sovereignty means the ultimate authority to make and enforce binding rules for a political community, over a defined territory and population, without subordination to external authority. The concept operates along two axes: internal sovereignty (final authority within the state) and external sovereignty (independence from foreign powers). Both have de jure and de facto dimensions, meaning that legal recognition of sovereignty matters, but so does practical capacity to enforce laws and protect assets in both the physical and digital realms (Pohle & Thiel, 2020).

In practice, sovereignty has never been absolute. States have always faced constraints from geography, economics, alliances, and the practical limits of enforcement. What emerging technology changes is not the existence of these constraints but their character. Control over digital infrastructure, data, and services has become constitutive of effective governance. When a state cannot secure its communications or its data, enforce its laws over digital activity within its borders, or access the computational resources necessary for economic competitiveness, its de facto sovereignty is being undermined.

Translating sovereignty to digital domains raises two key questions. First, over what digital activity does a state seek final authority? The answers must include data, infrastructure, platforms, algorithms, and computational capacity. Second, where does subordination arise through dependence on foreign states, foreign legal orders, or private platform power? A Canadian organization using an American cloud provider may find its data subject to American legal process even when stored on Canadian soil. A European government relying on American AI systems cannot be certain those systems will remain available during a geopolitical crisis. A Latin American country using a Chinese open-weight model cannot be certain what type of answers it may provide and how they may be contextualized.

"Digital sovereignty" entered mainstream policy discourse in the early-to-mid 2010s, accelerating after the Snowden revelations of 2013 exposed the extent of signals intelligence collection through commercial technology platforms (Pohle & Thiel, 2020). The term has intensified with U.S.-China competition, COVID supply chain shocks, and AI's emergence as a strategic capability. The related concepts of data sovereignty, technological sovereignty, and cyber sovereignty capture overlapping concerns, but "digital sovereignty" has emerged as the umbrella term for state interests in the digital domain (Fratini et al., 2024).

As the concept has gained prominence, its meaning has remained contested. This is why a structured framework is necessary.

Finally, a comprehensive understanding of sovereignty must also recognize that sovereignty considerations differ across orders of government. First Nations, Métis, and Inuit

governments are advancing distinct claims to AI and data sovereignty over their Peoples and territories. These claims are rooted in inherent rights and self-determination that predate and operate independently of federal and provincial frameworks. This paper does not attempt to address those distinctions in depth. Indigenous-led organizations better articulate their own perspectives and are doing this work directly, including the First Nations Information Governance Centre and Animikii. For further reading, there are recent perspectives published in Policy Options on AI and Indigenous data sovereignty (Tu, 2025) and by Natiea Vinson, CEO of the First Nations Technology Council (Vinson, 2025).

V.B Five Dimensions of Digital Sovereignty

Digital sovereignty is not a single thing that a nation either possesses or lacks. It is better understood as a composite of distinct dimensions, each concerning different objects, tested by different questions, and threatened by different vulnerabilities. The framework presented here identifies five dimensions that structure the analysis throughout this report.

Jurisdictional sovereignty asks whether a state can set and enforce binding digital rules without displacement by foreign legal orders. The primary threat is extraterritorial legal reach. For example, the U.S. CLOUD Act (2018) empowers American authorities to compel U.S.-headquartered companies to provide data from their servers regardless of where it is stored. FISA Section 702 provides even more substantive powers: unlike the CLOUD Act, which permits legal challenge, cloud service providers (CSPs) "are not able to disclose when they have been compelled under section 702" (Jarnecki, 2025). Jarnecki also notes that although FISA poses a valid concern, it requires meeting stringent criteria before it can be applied, and CSPs may push back against directives to disclose customer data. Additional practical challenges include users implementing strong encryption to safeguard their information, which provides an extra layer of protection. The May 2025 case of the International Criminal Court, where Microsoft blocked the Chief Prosecutor's email following U.S. sanctions, demonstrated that American legal reach extends beyond data access to service availability itself (Jarnecki, 2025).

Operational sovereignty concerns whether government and critical services can keep functioning securely under attack, outage, or geopolitical shock. The primary threats are cyberattacks, espionage, and illegal data exfiltration, i.e., nation-state and criminal actors targeting infrastructure through ransomware, supply chain attacks, persistent intrusion, or unauthorized extraction of sensitive data. Consider the 2024 CrowdStrike incident, a single faulty software update which impacted 8.5 million Windows hosts and created estimated business costs of $10 billion (Jarnecki, 2025). Or the historic Nortel case, where state-sponsored actors maintained access for nearly a decade, which contributed to the collapse of what was once Canada's most valuable technology company (CBC News, 2012).

Technological sovereignty asks whether, if a critical technology becomes unavailable or untrusted, a country can credibly migrate or replace it within a reasonable timeframe. The primary threat is vendor lock-in, i.e., technical and contractual barriers that prevent switching providers even when switching would be in the national interest. Lock-in takes multiple forms, including technical (proprietary systems, limited data portability), commercial (egress fees, contractual barriers), organizational (institutional inertia), and personnel (workforce trained on specific platforms). For example, cloud and AI markets are highly concentrated: AWS, Microsoft, and Google control approximately 63% of global cloud infrastructure spending (Synergy Research Group, 2025). Ukraine's deliberate design of its Delta system as "cloud agnostic," with engineers confident they could migrate to any provider with minimal lead time, demonstrates that lock-in is a design choice, not an inevitability (Jarnecki, 2025).

Societal sovereignty concerns the integrity of collective self-government in a digitally mediated public sphere. Can the political community form preferences and make decisions without systematic manipulation or opaque platform control? The primary threat is democratic distortion, where foreign interference, platform manipulation, and disinformation shapes public discourse in ways that a state cannot easily see or counter. AI-generated content, such as

synthetic media and deepfakes, amplify these risks and erode trust, while AI algorithms can propagate false or divisive content. Different approaches to model design, construction of training datasets and underlying governance rules create "not a single AI world, but overlapping cognitive infrastructures, each generating its own truths, exclusions, and forms of abstraction (Nguyen, 2025)." In short, societal sovereignty is about retaining the ability to collectively determine how society generates and processes what counts as knowledge itself.

Economic sovereignty asks whether domestic firms retain meaningful bargaining power and freedom to operate, or whether external actors can unilaterally impose terms, extract rents, or threaten cutoff. The primary threat is economic coercion through weaponized dependencies, including supply chain leverage wielded as a geopolitical tool. China's 2025 rare earth export controls, which now require approval for products containing even 0.1% Chinese-sourced materials, demonstrate how chokepoints in the supply chain can be weaponized (Szczepanski, 2025).

Coercion on digital sovereignty can also come through trade negotiations. Commerce Secretary Lutnick has explicitly demanded that the European Union ease tech rules, including the AI Act, Digital Services Act, and Digital Markets Act, in exchange for lower steel tariffs, framing European regulation as protectionist non-tariff barriers (Valero et al., 2025). This demonstrates economic coercion through weaponized regulatory pressure, where market access becomes contingent on regulatory concessions. Market concentration reduces bargaining power, resulting in higher prices, less favourable contract terms, and reduced ability to negotiate requirements.

Table V.1: Five Dimensions of Digital Sovereignty

Dimension	Core Question	Primary Threat
Jurisdictional	Can you enforce your rules?	Extraterritorial legal reach
Operational	Can you keep functioning under attack?	Cyber threats (attacks, espionage, exfiltration)
Technological	Can you migrate if needed? Can you substitute?	Vendor lock-in
Societal	Can you form and express preferences freely?	Democratic distortion, misinformation
Economic	Do you retain freedom to operate?	Economic coercion

V.C Canada's Challenge: Strategic Autonomy and Core Tensions

As a middle power with deep economic, technological, and security ties to the United States, Canada faces a distinctive challenge in approaching digital sovereignty.

Canada, like most countries, cannot achieve complete technological independence. The barriers are structural and daunting: the United States controls 75% of global AI compute capacity (Abecassis et al., 2025) and Taiwan Semiconductor Manufacturing Company manufactures 90% of advanced chips (Sastry et al., 2024). The cost of frontier AI training is projected to reach $6.6 billion per run by 2028, which likely exceeds what any single middle

power can sustain without wartime economic mobilisation (Abecassis et al., 2025). Complete self-sufficiency is neither achievable nor, given the pace of innovation, desirable.

The Government of Canada's Digital Sovereignty Framework admits to this challenge (Treasury Board of Canada Secretariat, 2025b). The framework defines digital sovereignty as "the ability to exercise autonomy over its digital infrastructure, data and intellectual property." Critically, it acknowledges that "It is impossible […] to obtain a state of complete digital sovereignty, known as digital autonomy, due to the absolute interconnected nature of the digital world."

Even within areas where having control is more realistic, pursuing digital sovereignty involves genuine trade-offs that must be acknowledged. Some security measures may limit access to best-in-class technology: the most capable AI models are developed by American labs, and restrictions on their use could constrain innovation. Building domestic alternatives also trades off against efficiency: sovereign infrastructure typically costs 15–20% more than hyperscaler alternatives (Maisto, 2025b). Accepting some sovereignty costs will likely be necessary, but the magnitude of those costs depends on how thoughtfully measures are designed.

Finally, trade-offs also clearly arise between digital sovereignty dimensions. Building domestic cloud infrastructure could address jurisdictional sovereignty concerns by keeping data under Canadian legal authority. However, if that infrastructure is not built to the highest security standards, it could create new operational sovereignty risks by providing a more vulnerable target for cyberattacks. A domestic AI model fine-tuned on Canadian data could enhance societal sovereignty, but if it performs significantly worse than foreign alternatives, the economic costs of using it could undermine economic sovereignty.

Pullquote here if needed for visual interest and entry point, pullquote. Dere prerib volup solupiet parita simi

The question, then, is not whether to have dependencies, but how to manage them, and how to do so within the constraints of Canada's existing legal and trade commitments.

Canada has operated within an international rules-based order. While we may be entering a new order, existing rules (Major, 2026) and agreements will likely still shape many aspects of trade, investment, and security relationships. CUSMA, WTO agreements, and bilateral arrangements constrain what sovereignty-affirming measures Canada can pursue. But these constraints are not purely external impositions: Canada depends on these same agreements for market access, dispute resolution, and predictable treatment of Canadian firms abroad. Unilaterally undermining trade rules to strengthen sovereignty would invite retaliation and erode the very frameworks that protect Canadian interests. The challenge is to pursue digital sovereignty aims within these constraints, using the policy space that trade agreements provide (particularly national security exceptions and government procurement carve-outs) rather than acting as if no constraints exist.

Any digital sovereignty strategy for Canada must therefore balance competing considerations: gains in one dimension may create risks in another, and measures should remain, when possible, defensible under international trade law.

V.D From Digital Sovereignty to AI Sovereignty

AI amplifies and extends many of the challenges posed by digital sovereignty. For example, when AI models process government data through foreign-controlled inference infrastructure, they potentially compound every sovereignty dimension simultaneously: jurisdictional exposure through CLOUD Act reach; operational risk through service dependency and exposure to stronger AI-powered cyberattacks; technological lock-in through provider-specific integrations; societal implications through model training choices and algorithmic decisions; and economic leverage through compute concentration.

Table V.2: AI Amplification of Sovereignty Dimensions

Dimension	Examples
Jurisdictional	AI services compound CLOUD Act exposure; inference requests are data flows subject to foreign legal process
Operational	AI systems present expanded attack surface; model weights are high-value intelligence targets
Technological	Model-specific APIs and fine-tuning investments create switching costs; training costs act as barriers to domestic alternatives
Societal	AI-generated content at scale; deepfakes; algorithmic manipulation of public discourse
Economic	Compute concentration creates acute dependency, allowing for value capture and extraction

AI also introduces considerations without direct parallel in prior digital sovereignty debates. For example, model weights represent a new category of sovereign asset: they are neither traditional data nor software, but encoded knowledge that can be exported, stolen, modified, or restricted. The distinction between training and inference creates different sovereignty postures: training infrastructure is highly concentrated and capital-intensive, while inference is more distributed and achievable for sovereign operations (Brookfield, 2025). Open-weight models can be retrained, audited, or adapted to local requirements in ways that proprietary models cannot Nguyen (2025). However, open-weight models alone often do not fully eliminate foreign dependency.

Moreover, as AI accelerates economic growth, scientific discovery, and national security capabilities, sovereignty can be threatened by not having access to the technology. A country that cannot participate in AI development or diffuse AI across its industry and society will become economically weaker, less competitive, and strategically vulnerable. An effective AI sovereignty strategy must therefore balance two imperatives: mitigating loss of sovereignty from foreign sources while building sufficient domestic capacity to benefit from the technology.

The path forward for middle powers like Canada lies in balancing these AI-specific risks to sovereignty against the potentially profound benefits the technology may bring. The goal is preserved choice and optionality, reduced leverage by any single partner, and, where possible, mutual accountability.

VI. REGULATORY AND TRADE CONTEXT

The previous section established AI sovereignty as a distinct set of goals deriving from the broader concept of digital sovereignty. But sovereignty also exists within a web of legal constraints. Canada both exercises sovereignty through domestic law and faces constraints through trade obligations. This section maps that terrain: the domestic legal tools Canada possesses, the extraterritorial legal threats it confronts, and the trade agreement constraints that shape available policy options. Understanding this landscape is essential for evaluating the strategic options developed in subsequent sections.

VI.A Canada's Legal Landscape

Canada has a fragmented regulatory framework for AI and data. Federal privacy law remains outdated with no clear AI law in place, and provinces have moved ahead in the absence of federal leadership. Canada's government has developed sophisticated AI governance for its own adoption while failing to extend it to the private sector.

The Personal Information Protection and Electronic Documents Act (PIPEDA, 2000) governs private-sector handling of personal information. PIPEDA's approach to cross-border transfers relies on accountability principles rather than explicit prohibitions. According to the Office of the Privacy Commissioner, organizations must use "contractual or other means" to provide comparable protection and must disclose that data may be processed in other countries and may be accessible to foreign authorities (OPC, 2009).

PIPEDA was designed pre-cloud and pre-AI, and modernization efforts have repeatedly stalled. AI Minister Evan Solomon has indicated that new privacy legislation is forthcoming, but that the AI and Data Act will not be reintroduced (Lunau, 2026; Walsh & Guilmain, 2026). Regulation is signalled to be "light, tight, and right," and shaped by over 10,000 consultation responses as well as the national AI Taskforce (Scott, 2025a; ISED, 2025d). The summary of the inputs received by the government as well as the individual taskforce reports and the full data of the online public consultation were posted online in February 2026 (ISED, 2026b).

On modernized privacy legislation, some provinces have filled the vacuum. Quebec's Law 25, fully in effect since September 2024, requires organizations using personal information for automated decisions to inform individuals, disclose the factors involved, and provide the right to request review and rectification of data held, with penalties reaching CAN$25 million or 4% of global revenue (Caron, 2024; Secure Privacy, 2024). British Columbia and Alberta maintain principles-based frameworks but lack specific AI provisions. The practical effect may be a Quebec effect: national companies will increasingly adopt Law 25 compliance as their baseline. The previously introduced Consumer Privacy Protection Act (under Bill C-27) had provisions to harmonize with provincial jurisdiction if the provincial privacy law was substantially similar (which Quebec's Law 25 was deemed to be) (David Young Law, 2023; Office of the Privacy Commissioner of Canada, 2023; Schaan, 2023).

The federal government has developed sophisticated AI governance for its own operations: the Treasury Board Directive on Automated Decision-Making requires Algorithmic Impact Assessments, risk classification, and human oversight; the Direction for Electronic Data Residency requires protected government data to remain in Canada (Treasury Board of Canada Secretariat, 2025a; Treasury Board of Canada Secretariat, 2018).

VI.B Cross-Border Legal Exposure

Canadian laws and regulations are not the only legal frameworks that apply to data and AI services consumed in Canada.

The Core Problem: Data Residency Does Not Equal Control

A common misconception holds that storing data in Canada provides legal protection. The reality is more troubling: if the provider is subject to foreign law, data is exposed to foreign legal processes regardless of physical location. This affects any service from foreign-headquartered companies, including cloud services, foundation model APIs, and Software as a Service applications.

The Government of Canada has stated the position clearly: "As long as a CSP that operates in Canada is subject to the laws of a foreign country, Canada will not have full sovereignty over its data. This is because there remains a risk that data stored in the cloud could be accessed by another country" (Treasury Board of Canada Secretariat, 2018). The Treasury Board's 2025 Digital Sovereignty Framework reiterates: "Using a Canadian supplier or storing data in Canada does not guarantee data will be outside the jurisdiction of foreign courts" (Treasury Board of Canada Secretariat, 2025b). The government can only maintain full legal control when it delivers services itself or uses providers operating entirely under Canadian jurisdiction.

U.S. Extraterritorial Legal Reach

The U.S. CLOUD Act (2018) compels U.S. companies to produce data in their "possession, custody, or control" regardless of where it is stored. The mechanism is relatively straightforward: a U.S. legal order is served on a company subject to U.S. law, which must then produce the data unless it takes legal action. The scope of Canadian exposure is substantial: it is estimated that at least one-third of the near-300 data centres in Canada are American-owned (Kasonga, 2025). The federal government's approved cloud provider list includes seven U.S. companies and just one Canadian (McLauchlan, 2025).

Major cloud service providers have diminished the significance of the CLOUD Act. For example, AWS' website says: "When we receive such requests for enterprise customer content, we make every reasonable effort to redirect law enforcement to the customer and notify the customer when legally permitted. If we are required to disclose customer content, we notify customers before disclosure to provide them an opportunity to seek protection from disclosure unless prohibited by law" (Kimball, 2025). The act also preserves standard comity protections and judicial review requirements. AWS claims zero disclosures of enterprise or government content stored outside the United States since it began tracking in 2020 (AWS, 2026). And the Government of Canada claims "no documented cases of foreign governments seeking access to the data of Canadian enterprises held by suppliers" (Treasury Board of Canada Secretariat, 2025b).

These facts and historical data should inform risk assessment, but do not fully eliminate concern. Some of the above facts come from interested parties. They also do not address consumer accounts, challenged requests, or National Security Letters subject to gag orders.

Further, beyond the CLOUD Act, it is worth noting that FISA Section 702 (50 U.S.C. § 1881a) permits surveillance of non-U.S. persons abroad with the compelled assistance of electronic communications service providers, and directives issued under Section 702 include nondisclosure requirements that prohibit providers from revealing the existence of surveillance orders or their compliance (Brennan Center, 2026). Separate criminal penalties of up to eight years' imprisonment apply to the unauthorized disclosure of classified information acquired under these authorities (50 U.S.C. § 1881h).

The difference between how U.S. and Canadian legal regimes treat privacy rights is also consequential for data sovereignty. Section 8 in the Canadian Charter of Rights and Freedoms creates a strong expectation of privacy in digital information, generally requiring judicial authorization for state access (R. v. Spencer, 2014). Canadian courts have consistently held that sharing information with a service provider does not diminish an individual's privacy expectations. This stands in contrast to the U.S. "third-party doctrine" that reduces privacy protections when data is held by intermediaries (R. v. Spencer, 2014). This divergence means Canadian privacy expectations for data held by cloud providers may have some incompatibilities

with U.S. legal frameworks that treat such data as more readily accessible. The Citizen Lab observes, "one would be hard pressed to find two democracies that are more incompatible when it comes to trying to align digital surveillance laws" (Khoo & Robertson, 2025).

Overall, while the U.S. remains a partner of Canada on many fronts, in the context of AI sovereignty it is worth being mindful of the differences in jurisdictional frameworks, principles, and expectations on digital privacy between the two countries.

VI.C Trade Agreement Constraints

Canada has signed multiple trade agreements constraining digital sovereignty measures. Understanding these constraints, and the policy space that remains, is essential for developing viable strategic options.

CUSMA Chapter 19: Key Constraints

CUSMA Chapter 19 on Digital Trade, effective July 1, 2020 (Global Affairs Canada, 2020), established rules with no equivalent in NAFTA. The constraints are significant:

»» Article 19.4 (Non-Discrimination): Requires parties not to treat foreign digital products less favorably than domestic products. U.S. industry has already cited this provision in challenging Canadian measures including the Online News Act (CCIA, 2022), while the Digital Services Tax dispute focused on services and investment obligations under Articles 15.3 and 14.4 (USTR, 2024).

»» Article 19.11 (Cross-Border Data Flows): Prohibits restricting cross-border transfer of information for business conduct. An exception exists for "legitimate public policy objectives," but is subject to strict necessity tests.

»» Article 19.12 (Data Localisation): Prohibits requiring companies to use or locate computing facilities in a local territory as a condition for conducting business. Unlike CPTPP, CUSMA does not allow a "legitimate public policy objective" exception for this provision (Leblond, 2019).

»» Article 19.16 (Source Code): Cannot require foreign companies to disclose algorithms as a market entry condition, though regulatory and judicial access for specific proceedings is permitted.

These provisions substantially constrain Canada's options for commercial data governance. Data residency requirements for private-sector operations, mandates for local computing infrastructure, and policies that prioritize Canadian digital service providers all face potential challenges. According to Keith Jansa, CEO of the Digital Governance Council, the effect is a "structural asymmetry." Jansa remarks that "CUSMA is written as reciprocal obligations on paper, but reciprocity does not mean the gains are evenly shared in practice. Digital rules […] naturally reward the players already operating at scale across the border, which, in North America, are disproportionately U.S.-based platforms and cloud firms" (Jansa, 2025).

Government Procurement: A Different Regime

Critically, CUSMA Chapter 19 does not apply to government procurement (Global Affairs Canada, 2020, art. 19.2(3)). This creates a two-track regime: while the commercial market is heavily constrained, government procurement is not bound by Chapter 19's digital trade disciplines. Canada-U.S. government procurement is instead governed by the WTO Government Procurement Agreement (GPA), which requires non-discrimination for covered procurement above approximately CAN$630,000 for services (Global Affairs Canada, 2025).

These different rules allow for policy space to exist within this framework. Outcome-based technical specifications, requiring certain security certifications or jurisdictional controls rather than Canadian ownership, may be defensible. Below-threshold procurement and non-covered entities retain flexibility. The National Security Exception (NSE) (Global Affairs Canada, 2022, art. 32.2) is self-judging, allowing measures Canada "considers necessary" for protecting its own essential security interests (Office of the Procurement Ombud, 2025).

Canada has already exercised the NSE for sovereign AI. Shared Services Canada's 2025 Sovereign Public Cloud procurement invokes the NSE to exempt the procurement from the WTO GPA and Canadian Free Trade Agreement,

enabling requirements that data be processed exclusively in Canada by providers not subject to foreign access laws (Shared Services Canada, 2025a). This demonstrates a viable legal pathway, one that can inform broader strategic options.

The Evolving "Buy Canadian" Landscape

The Government of Canada unveiled a Buy Canadian Policy in Budget 2025, with further details unveiled on December 16, 2025 (PSPC, 2025a). Information and communications technology is one of five strategic sectors where Canadian suppliers receive evaluation advantages through bid price discounts and scoring for Canadian content. For government cloud procurement, this creates opportunity for proponents of domestic/sovereign cloud services. The policy faces criticism from multiple directions: U.S. industry groups warn it could "result in a more fragmented market that also denies Canadian government customers access to leading-edge technologies" (CCIA, 2025a). On the other hand, Canadian analysts argue the "Canadian supplier" definition allows foreign multinationals to qualify through local subsidiaries (O'Toole & Pence, 2026). Some critics remark that the new definition may still be too vague to solve "the origin loophole" entirely, where companies rely on a Canadian mailbox and non-Canadian personnel (Andrews, 2025). Amendments preventing challenges to "Buy Canadian" policies before the Canadian International Trade Tribunal raise additional trade compliance questions. With thresholds dropping from CAN$25 million to $5 million by spring 2026 (PSPC, 2025b) and reciprocal procurement policies fully launching in spring 2026, Canadian procurement rules remain a rapidly developing policy space.

VI.D The 2026 CUSMA Review

CUSMA includes a scheduled review mechanism. The first joint review is due July 1, 2026, the sixth anniversary of entry into force. If the parties do not agree to extend the agreement, it will terminate in 2036 (Bitar et al., 2025). However, the more severe risk is that the United States triggers a clause that would allow them to withdraw with only six months' notice. Regarding CUSMA, President Donald Trump has stated that "it wouldn't matter to me" if the agreement expired, while U.S. Trade Representative Jamieson Greer has confirmed that withdrawal is "always a scenario" and "all of those things are on the table" (Malone, 2026; Palmer, 2025; Crawley, 2025).

Within this context, U.S. industry is positioning aggressively. The Computer & Communications Industry Association's (CCIA) November 2025 submission urged preserving CUSMA's digital trade provisions, highlighted "discriminatory" Canadian measures including the Online News Act and Online Streaming Act, and noted U.S. digital exports to Canada and Mexico exceeded $75 billion in 2024 (CCIA, 2025b). The U.S. Chamber of Commerce demanded "full repeal" of the Online Streaming Act (Malone, 2025). USTR Greer has signalled an aggressive digital trade enforcement agenda. The CCIA remarks welcoming his appointment note that Greer is "standing up against unreasonable or discriminatory barriers, including in the digital space" (Curl, 2025; CCIA, 2025c). Canada's June 2025 rescission of the Digital Services Tax (Department of Finance Canada, 2025) to resume trade negotiations illustrates the pressure Canada is under.

The review represents both risk and opportunity. U.S. priorities will likely include further constraining Canadian digital sovereignty measures, locking in Chapter 19 provisions, limiting use of security exceptions, and restricting "Buy Canadian" policies in the tech sector. Canada must be prepared to defend existing policy space while the strategic framework developed in subsequent sections informs how Canada could approach these negotiations.

VII. ASSESSING CANADA'S SOVEREIGN AI CAPACITY

The previous sections established the types of sovereignty threats and the legal and trade landscape that Canada faces. This section applies this knowledge by systematically addressing each layer of Canada's AI technology stack, examining Canada's position, identifying key actors, assessing vulnerability across the five dimensions of digital sovereignty, and identifying potential areas for action. (See Appendix A for detailed layer definitions.)

Layer	Description
1. Data and Data Governance	Training datasets, operational data, and governance frameworks
2. Physical Infrastructure and Networks	Data centres, telecommunications, and energy systems
3. Compute Hardware	Processors, memory, and networking equipment enabling AI workloads
4. Cloud Infrastructure Services	Virtualized compute, storage, and managed services (IaaS/PaaS)
5. Foundation Models (Inference)	Access to and deployment of large-scale AI models
6. Model Operations and Orchestration	Tools for deploying, serving, and monitoring AI in production
7. Applications	End-user AI products and AI-enabled software (SaaS)

This assessment focuses specifically on the AI technology stack for deploying and running AI systems (inference), rather than for training frontier foundation models or providing general-purpose computing infrastructure. This focus reflects three strategic realities. First, AI inference is expected to account for 75% of total AI compute demand by 2030, representing where the majority of new compute investment will occur (Brookfield, 2025). Second, inference aligns with the economic and productivity benefits that justify sovereign AI investment: this is AI being used in production. Third, AI training has fundamentally different requirements and economics, warranting separate treatment in Section IX.

Scope: Inference vs. Training Infrastructure

This assessment focuses on the AI stack for deploying and running AI systems (inference), not for training frontier foundation models. Training infrastructure has fundamentally different requirements.

Aspect	Inference/Deployment	Training
Compute scale	Distributed, ongoing	Massive, concentrated
Hardware	Inference-optimized, flexible	Training-specific
Power	Standard data centre	Uninterruptible for weeks or months
Data	Retrieval-focused, operational data	Petabyte-scale training datasets
Tooling	Model serving, orchestration	Distributed training frameworks

As of mid-2025, only 13 OECD countries hosted public cloud compute relevant to training frontier models, with the United States the only country offering multiple H100-equipped regions. While new investments announced in Canada, the U.K., Germany, China, and elsewhere may expand this number, the underlying concentration of training-grade compute remains mostly within the U.S. (Lehdonvirta et al., 2025).

VII.A Data and Data Governance (Layer 1)

Data is the foundational input to the AI stack, the raw material from which models learn and the contextual information that makes AI systems useful. This layer encompasses training data, fine-tuning data, operational data accessed via retrieval systems, personal and user data, and the governance frameworks that determine who controls access to these strategic assets.

Data serves multiple functions across the AI lifecycle. Training data comprises the petabyte-scale datasets from which foundation models learn patterns of language, reasoning, and knowledge. Common Crawl, an open repository of web data, contains over 9.5 petabytes of web data collected since 2008. It provided more than 80% of GPT-3's training tokens and was used by two-thirds of large language models released between 2019 and 2023 (Mozilla Foundation, 2024). Fine-tuning data consists of domain-specific datasets that adapt general-purpose models to specialized tasks. Operational data, accessed via retrieval-augmented generation, provides the real-time information that makes AI systems useful in production. User and personal data are key inputs to personalizing model behaviour and outputs. Governance frameworks constitute the policy layer determining who controls access to these strategic assets.

Canada possesses valuable domain-specific data assets that represent genuine strategic resources. For example, CanLII provides free access to Canadian court judgments and legislation with over 99,000 AI-generated bilingual summaries (CanLII, 2025). The Translation Bureau maintains eight billion words

of bilingual government text, which provides a basis for GCtranslate, an AI prototype for official languages translation for federal public servants (PSPC, 2025c). CIHI holds comprehensive health data; OSFI maintains detailed financial data including mortgages and consumer lending; Statistics Canada and the Canada Revenue Agency hold extensive administrative datasets. However, commercially, a paper by OpenText cites that an estimated 90% of valuable data resides behind enterprise firewalls in healthcare systems, financial institutions, telecommunications networks, and industrial operations, and is largely inaccessible for sovereign AI development (Bell, Fraser & Jenkins, 2025; Davis, 2019). The challenge is not ownership but how to leverage Canada's strategic data assets to generate major value, given the lack of coordination mechanisms or data use strategies.

Societal sovereignty is the primary concern at the data layer. Canadian content is significantly underrepresented in the training datasets that shape AI systems. Common Crawl's dominance in LLM training means that whatever it captures, and whatever it excludes, shapes the cultural assumptions embedded in AI. Quebec French accounts for less than approximately 5% of French-language recordings in the Common Voice dataset. Models fine-tuned on European French perform worse on Quebec French than original multilingual versions (Serrand, Boulianne & Morsli, 2025). When the datasets that shape AI systems systematically underrepresent Canadian content, the resulting models embed foreign assumptions about language, culture, and norms. AI models are "cultural products" embedding creator biases (Escott, 2025). For a bilingual, multicultural country, this matters: AI systems trained predominantly on American English and European French encode different legal concepts, historical narratives, and social norms than those reflecting Canadian contexts.

Jurisdictional and operational sovereignty are secondary risks for Canada at this layer. The actual vulnerabilities are primarily in the physical and cloud infrastructure layers, as well as at the application player. Layer 1's role is preventative: data governance frameworks determine what data can flow where, under what conditions, and whether sensitive information reaches jurisdictionally compromised infrastructure. Strong data classification and handling policies reduce exposure at those vulnerable layers. The current regulatory landscape is poorly positioned for this role, with outdated privacy laws and no AI legislation at the federal level.

Vulnerability assessment: Moderate. Canada has valuable data assets but lacks mechanisms to leverage data for sovereign AI development, while Canadian content remains underrepresented in global training datasets. Further, while Canada at least has a federal data privacy law in PIPEDA, as of early 2026, Canada does not have a national data strategy and does not have modern data privacy laws (updated within the last five years) across most of the country.

Strategic options: Data classification policies, government data governance frameworks that enable AI development while maintaining sovereignty, incentives for Canadian content creation in training datasets, and coordination with data platforms and Canadian content creators to improve representation in global datasets.

VII.B Physical Infrastructure and Networks

Physical infrastructure comprises the tangible assets that power AI systems: data centres, telecommunications networks, and the energy systems that sustain continuous operation. Canada has abundant clean energy, relatively low electricity prices in some regions, and a cool climate. These are all factors that can help attract substantial investment, which positions Canada as one of the world's best hubs for digital infrastructure (Canada Energy Regulator, 2024).

The Canadian data centre market reached US$5.44 billion in 2024 and is projected to reach US$12.27 billion by 2030 at 14.5% compound annual growth (Arizton, 2025). IT load capacity stands at 3.13 GW (Mordor Intelligence, 2025) with over ten GW in additional capacity expected from upcoming colocation and data-centre projects (ResearchAndMarkets, 2026). Roughly 80% of Canadian electricity is non-emitting and 67% is renewable (Canadian Centre for Energy Information, 2026).

Canadian-owned operators provide domestic alternatives to hyperscalers. eStruxture Data

Centers, headquartered in Montreal, operates 15 facilities and frames itself as Canada's largest homegrown data centre provider (eStruxture, 2026). QScale in Levis, Quebec offers 142 MW of secured capacity supported by 100% renewable energy (QScale, 2026). Bell Canada announced an AI Fabric partnership targeting up to 500 MW across six facilities (Bell Canada, 2025b), while TELUS opened a Sovereign AI Factory in Rimouski, Quebec in September 2025 (TELUS, 2025).

Telesat's Lightspeed LEO constellation, comprised of 156 satellites with the full constellation scheduled for launch in 2027, will provide Canadian-owned connectivity options (Katz, 2025; Rainbow, 2025). While satellites and space technology are not explored in-depth in this report, it should be noted that they may be a point of vulnerability given that an increasing amount of internet communication is happening via satellite connections. Further, many geospatial satellites were recently found by researchers (primarily from the University of California San Diego) to be broadcasting unencrypted information, which is a potential security vulnerability for Canadian data (Zhang et al., 2025).

American firms own almost one-third of Canada's nearly 300 data centres (Kasonga, 2025) and hyperscalers dominate capacity expansion. AWS has committed CAN$24.8 billion through 2037 (AWS, 2023) and Microsoft recently announced CAN$7.5 billion in new data centre spending in Canada over two years (Smith, 2025). The distinction between physical infrastructure ownership and cloud services operation can often be blurred: major hyperscalers are vertically integrated, owning and building their own data centres while simultaneously providing cloud compute services on top of that infrastructure. This vertical integration means that sovereignty concerns at Layer 2 (physical infrastructure) and Layer 4 (cloud services) often involve the same corporate actors.

Jurisdictional sovereignty is the primary concern at this layer, arising from network routing vulnerabilities. More than 25% of Canadian domestic internet traffic routes through the United States. Traffic routed through the U.S. is subject to FISA and the U.S Patriot Act surveillance authorities (Clement & Obar, 2014). The Canadian Internet Registration Authority has supported the expansion of domestic Internet Exchange Points significantly (CIRA, 2026), but the routing issue persists. This is jurisdictional exposure through infrastructure design rather than contractual relationship.

Vulnerability assessment: Low-moderate. Canada's energy advantage has attracted substantial investment and Canadian-owned operators provide domestic alternatives. Network routing through U.S. jurisdiction represents an underappreciated vulnerability requiring further assessment.

Strategic options: Direct investments in Canadian-owned data centre capacity, network routing optimization through domestic Internet Exchange Points, end-to-end encryption for sensitive traffic traversing foreign infrastructure, and regulatory incentives for sovereign infrastructure development.

VII.C Compute Hardware

Compute hardware is comprised of the processors, memory systems, and networking equipment that enable AI workloads. Chip manufacturing is arguably the most complex supply chain in the world, requiring thousands of specialized inputs, decades of accumulated process knowledge, and fabrication facilities costing tens of billions of dollars. This layer represents the most concentrated element of the global AI stack. It is Canada's most severe structural vulnerability, one that cannot be meaningfully resolved through domestic industrial policy alone.

NVIDIA, headquartered in the United States, dominated 92% of the GPU market in 2025 and holds over 80% of the AI hardware market (Carbon Credits, 2026). The company's CUDA software ecosystem creates deep lock-in with over four million developers (Tran, 2026). Nascent alternatives to NVIDIA exist, including AMD's ROCm ecosystem (also U.S.-based), Google's TPUs, and RISC-V architectures. However, Google's Tensor Processing Units (TPUs) deserve increasing attention. While historically available only through Google Cloud, Google has reportedly begun offering TPU access to third parties, potentially diversifying the accelerator landscape (Butler, 2025; Chen, 2025; Chernicoff,

2025). And TPUs are powering Gemini 3, largely recognized as one of the most powerful models in the world. None of the leading AI chip companies are Canadian.

> Pullquote here if needed for visual interest and entry point, pullquote. Dere prerib volup solupiet parita simi

Foreign supply chain dominance extends upstream to fabrication and memory. Taiwan Semiconductor Manufacturing Company (TSMC), based in Taiwan, holds 72% of overall foundry market share and an estimated 90%+ of advanced logic chip fabrication specifically (Lai & Li, 2025; Kitishian, 2025). Samsung of South Korea provides the only alternative at scale. ASML of the Netherlands is the sole supplier of extreme ultraviolet lithography machines required for leading-edge chips, creating the ultimate chokepoint. Memory represents an additional bottleneck: as of 2022, SK Hynix of South Korea produced approximately 50% of high-bandwidth memory essential for AI accelerators, Samsung produced approximately 40%, and Micron of the United States produced approximately 10% (Liu & Ju, 2023). AI is also "skyrocketing" the price of RAM, which will have the secondary effect of a price increase on consumer goods (Ali, 2025). Long lead times reflect scarcity: data centre GPUs required 36 to 52 weeks from order to delivery in late 2023 (Galabov et al., 2023), and new fabrication capacity requires three to five years from investment to production (Intel, 2023).

Canada has some promising semiconductor companies, but they occupy niche segments of the technology stack and are not substitutes for the GPUs, high-bandwidth memory, and advanced fabrication capabilities that dominate AI compute. Tenstorrent, led by prominent chip architect Jim Keller, has raised more than $1 billion to pursue a dual strategy of building AI hardware systems and licensing IP, but quietly relocated its headquarters to the U.S. (Romburgh, 2024; Scott, 2024). Other emerging Canadian companies include RANOVUS and Talaas, which has recently announced a major new hardware-based inference product (Freund, 2026).

Hypertec of Montreal is the only Canadian-headquartered NVIDIA original equipment manufacturer. IBM's Bromont facility in Quebec is the largest Outsourced Semiconductor Assembly and Test facility in North America (Murphy, 2024) but is not a fabrication plant; the Canadian Photonics Fabrication Centre in Ottawa provides photonics prototyping but focuses on photonics rather than advanced logic chips. Canada is not included in the Chip 4 Alliance (the U.S., Taiwan, South Korea, and Japan) and has no CHIPS Act equivalent. The U.S. CHIPS Act allocated US$52 billion for semiconductor manufacturing and research (NIST, 2025); the EU Chips Act committed 43 billion euros (European Commission, 2024); Japan invested US$13 billion (Mochizuki, 2023). The Canadian semiconductor industry is comprised of over 500 companies but holds less than 1% of global market share (ISED, 2024; Joshi, 2025; Riehl, 2025).

Technological sovereignty is fundamentally compromised at this layer, as it is not feasible to meaningfully produce domestic substitutes at scale. Building sovereign chip manufacturing capability would require not only fabrication facilities costing US$20 billion or more (Shilov, 2025), but also decades of accumulated process knowledge, thousands of specialized suppliers, and the scale to justify such capital deployment. The ecosystem requirements extend beyond manufacturing to include chip design expertise, advanced packaging capabilities, materials science, and equipment suppliers; a value chain that no middle power can realistically assemble independently. Canada's chip industry has contributed and will continue to contribute meaningfully to the global chip and hardware supply chain through innovations in areas like photonics, advanced packaging, and specialized chip design. However, it has never comprised a large enough share of the overall supply chain to approach technological self-sufficiency in the components that matter most for AI: the design and manufacturing of GPUs, high-bandwidth memory, and leading-edge chip fabrication.

Economic sovereignty is acutely threatened at this layer. The supply chain concentration creates potential for coercion and disruption. During global shortages, allocation decisions made by NVIDIA and hyperscalers could shift away from Canadian customers toward those with greater strategic priority. Canada's exemption from U.S. export controls depends on maintaining a healthy bilateral relationship.

Five Eyes membership would typically provide preferential treatment, status as a trusted ally ensures access to controlled technologies, and defence cooperation secures priority allocation. In January 2025, the Biden administration published a three-tier export control framework for AI chips that classified 18 close allies including Canada in the top tier with favourable access through license exceptions (Bureau of Industry and Security, 2025a). However, the incoming Trump administration rescinded the framework in May 2025 before it took effect, calling it "overly bureaucratic" (BIS, 2025b; Sokler et al., 2025). As of early 2026, no replacement framework has been finalized, leaving allied access governed by earlier, less structured export control rules. This regulatory uncertainty itself illustrates the vulnerability: Canada's access to advanced AI chips likely depends not on binding agreements but on the policy preferences of the U.S. government.

This layer for AI chips fails both the technological sovereignty test ("can you substitute if needed?") and the economic sovereignty test ("do you retain freedom to operate?").

Vulnerability assessment: Critical. And these vulnerabilities cannot be resolved through domestic action alone.

Strategic options: Supply chain diversification across allied nations, strategic partnerships and participation in multilateral semiconductor initiatives, support for Canadian chip design companies, and contingency stockpiling of critical components. Large-scale domestic fabrication is not a realistic option.

VII.D Cloud Infrastructure Services

Cloud infrastructure services represent an acute but controllable vulnerability in Canada's AI technology stack; consequently, it is a strategic intervention point. Unlike AI compute hardware production, where NVIDIA and TSMC concentration cannot be resolved through domestic action, sovereign cloud infrastructure is achievable. Canada has some domestic alternatives, government procurement leverage, and viable legal pathways.

Cloud infrastructure services, commonly called Infrastructure as a Service (IaaS) or Platform as a Service (PaaS), provide virtualized compute, storage, and networking on demand. For AI specifically, this means access to GPUs and other accelerators as a service, enabling organizations to run AI workloads without owning hardware. The shift from on-premises servers to cloud computing represents one of the most consequential infrastructure transitions in decades, and AI is accelerating further cloud construction and use dramatically. By 2029, Gartner projects that 50% of cloud compute resources will be devoted to AI workloads, up from less than 10% today (Gartner, 2025a).

The Canadian cloud market reached an estimated US$54.78 billion in 2025 and is projected to reach US$140.75 billion by 2031 at 17.02% compound annual growth (Mordor Intelligence, 2026). Canada is experiencing particularly rapid growth: International Data Corporation reports that Canada showed 151.8% year-over-year growth in cloud infrastructure spending in Q4 2024. This is among the fastest rates globally, exceeding the United States at 125.3% (IDC, 2025).

Global hyperscalers dominate the market: AWS (29%), Microsoft Azure (20%), and Google Cloud (13%) control approximately 63% of enterprise spending on cloud infrastructure (Casey, 2025). The AI-specific cloud segment is growing even faster: Gartner reports AI-optimized IaaS (GPU as a service, AI accelerators) reached $18.3 billion in 2025 at 146% year-over-year growth, projected to hit $37.5 billion in 2026 (Gartner, 2025b). Globally, Gartner forecasts that worldwide public cloud spending will reach $723.4 billion in 2025, representing 21.5% growth from $595.7 billion in 2024 (Gartner, 2024).

The Government of Canada recognized cloud computing's strategic importance early. In 2018, the federal government introduced a "cloud-first" adoption strategy and launched protected-cloud procurement, requiring cloud-first consideration for new applications (Government of Canada, 2018). In 2023, the strategy was shifted to be "cloud smart" and updated its principles, including a move towards zero-trust architecture, "buy before build," and increased value and reduced technical debt (Government of Canada, 2023). Yet the resulting procurement framework reveals the persistent sovereignty challenge: as

of 2025, seven of eight cloud service providers on Shared Services Canada's (SSC) approved supplier list are American (Riehl, 2025; Shared Services Canada, 2025c). ThinkOn is the only Canadian company certified to sell cloud services to the federal government for Protected B workloads.

Canadian companies are responding to these needs with sovereign cloud initiatives that represent different trade-offs. The Sovereign Cloud Consortium (ThinkOn, Hypertec, Aptum, eStruxture), announced in October 2025, emphasizes end-to-end Canadian ownership which they claim will ensure jurisdictional separation from U.S. law (ThinkOn, 2025). Bell AI Fabric, announced in May 2025, maintains Canadian data residency and operational control, and partners with Cohere for sovereign AI deployment (Bell Canada, 2025b). TELUS opened Canada's first "Sovereign AI Factory" in Rimouski, Quebec in September 2025, stating they are the first North American service provider to join the NVIDIA Cloud Partner network (TELUS, 2025).

Jurisdictional sovereignty is the primary threat. As detailed in Sections V and VI, the U.S. CLOUD Act (2018) empowers American authorities to compel U.S.-headquartered companies to produce data in their "possession, custody, or control" regardless of where it is stored. Data residency does not equal data sovereignty when the provider is jurisdictionally subordinate to a foreign legal order. The SSC Sovereign Cloud Services Request for Information (2025) explicitly invokes the National Security Exception, demonstrating that the government has identified both the vulnerability and a viable legal pathway to address it (Shared Services Canada, 2025b).

Operational sovereignty is threatened through two vectors. First, service denial risk: a foreign government could force a cloud provider to unilaterally deny service to Canadian entities, as demonstrated by the ICC incident, in which Microsoft blocked access to ICC Chief Prosecutor Karim Khan's email following U.S. sanctions. This represents the "kill switch" vulnerability identified in Section III. Second, hyperscaler infrastructure presents high-value targets for data exfiltration through espionage. Concentration of sensitive data in cloud environments creates targets for intelligence collection outside of legal frameworks.

Technological sovereignty is compromised through documented vendor lock-in. Cloud concentration creates documented operational risks: 62% of organizations cite concentration as a "top five" risk (Gartner, 2023). When migration is practically impossible, the technological sovereignty test ("can you migrate if needed?") fails. Ukraine's deliberate design of its Delta system as "cloud agnostic," with engineers confident they could migrate to any provider with minimal lead time, demonstrates that lock-in is a design choice, not an inevitability (Jarnecki, 2025).

Economic sovereignty is threatened through market concentration and value extraction, with cloud compute dominated by Amazon, Microsoft, and Google. This concentration reduces Canadian bargaining power, enables rent extraction, and creates dependence. A recent study estimates EU companies spend 264 billion euros annually on American cloud and software services, with substantial sovereignty and economic implications, which is a cautionary model for Canada (Cigref, 2025).

Federal Compute Consortia Call

The federal government has moved beyond procurement frameworks to actively soliciting large-scale sovereign AI infrastructure projects. In January 2026, Innovation, Science and Economic Development Canada (ISED) launched a one-month initiative to identify promising AI infrastructure projects for potential memoranda of understanding. This is a direct follow-on from Budget 2025, which committed the government to explore mechanisms enabling large-scale commercial AI data centres (ISED, 2026a).

The initiative requests information on sovereignty conditions: projects are asked to share details about Canadian control, with the intake form directly asking whether the project will be "majority controlled by a Canadian-controlled organization" and whether "all of the data processed by your project [will] be domiciled in Canada." Preference appears to be given to proposals exceeding 100 MW that demonstrate clear paths to completion, Indigenous participation, environmental sustainability, and maximized use of Canadian partners and supply chains.

Critically, applicants must describe "how your project aligns with the Government of Canada's objective to create sovereign AI infrastructure," including ownership details, control structures, and intended clients. The initiative explicitly connects to existing federal mechanisms including the Sovereign Compute Infrastructure Program and Canada Infrastructure Bank. This is a strong signal that the government views sovereign cloud infrastructure as requiring coordinated industrial policy rather than market forces alone.

Vulnerability assessment: Critical. This layer concentrates threats across four sovereignty dimensions which makes it among the most vulnerable points in the stack. Vulnerabilities include U.S. CLOUD Act exposure despite Canadian data residency, service denial risk demonstrated by the ICC incident, vendor lock-in creating migration barriers, and economic sovereignty concerns from hyperscaler concentration.

Strategic options: Government procurement mandates for sovereign cloud infrastructure, direct support for domestic sovereign cloud providers, security certification frameworks that enable Canadian alternatives, and leveraging the National Security Exception for procurement requirements.

VII.E Foundation Models (Inference)

Foundation models are the critical suppliers of "intelligence" in the AI stack. Unlike traditional software where functionality is determined by code, foundation models learn capabilities from data and can be applied across a wide variety of tasks without task-specific programming. This makes them uniquely strategic: whoever controls access to capable foundation models controls a considerable portion of the value that the modern AI stack produces.

Three American companies dominate the global foundation model market. As of mid-2025, Anthropic (Claude) holds 40% of enterprise LLM spending; OpenAI (ChatGPT) holds 27%; and Google (Gemini) holds 21% (Tully et al., 2025). Combined, these three companies account for 88% of enterprise foundation model usage. In consumer markets, ChatGPT maintains 68–75% market share with over 800 million weekly active users. 92% of Fortune 500 companies have employees using ChatGPT (OpenAI, 2025a; Shum, 2026).

Canadian adoption mirrors these global patterns: 20% of online Canadians use ChatGPT, and 66% have experimented with generative AI tools (Summerfield, 2024; Gruzd et al., 2025).

Only 12.2% of Canadian businesses formally use AI, but adoption doubled from 6.1% in Q2 of 2024 according to Statistics Canada data (Bryan et al., 2025).

Cohere, headquartered in Toronto, is Canada's only domestic foundation model company and is an important hedge against complete foreign dependency. Its strategic differentiation lies in deployment flexibility: unlike major American competitors whose models are tightly coupled to specific hyperscaler platforms, Cohere maintains cloud-agnostic deployment and derives approximately 85% of its revenue from private deployments rather than API services (Asplund & Ballve, 2025). Its models can run on minimal hardware, enabling on-premises and air-gapped installations that meet stringent sovereignty requirements (TechCrunch, 2025). The Government of Canada signed a Memorandum of Understanding with Cohere in August 2025 recognizing it as a "strategically important LLM" (ISED, 2025e), alongside CAN$240 million in public funding under the Sovereign AI Compute Strategy (Department of Finance Canada, 2024b). Cohere has secured contracts with the Communications Security Establishment (Hemmadi, 2025), as well as partnerships with Bell Canada for sovereign deployment (Bell Canada, 2025a) and Thales for defence applications (Thales, 2025).

Open-source and open-weight models provide an additional strategic hedge against foreign model dependency. Meta's Llama 4 Maverick and Mistral's Large 3 achieve 85–90% of frontier model performance at significantly lower cost, or zero cost (beyond compute) for self-hosted deployments (Digital Applied, 2025). However, not all open models are suitable for sovereign deployment. Canada banned DeepSeek on some federal government devices (Deschamps, 2025) due to concerns about data transmission to China. Both the U.S. National Institute of Standards and Technology and independent security researchers have identified DeepSeek models as significantly more vulnerable to attack than American alternatives (NIST, 2025).

Economic sovereignty is the primary concern at this layer due to single Canadian vendor concentration. Cohere is Canada's only credible foundation model developer: if Cohere fails or is acquired by a foreign company, Canada loses its primary hedge against AI model dependency on American providers. The challenge is not that Cohere's models are uncompetitive. Rather, it is impossible to compete across all dimensions on which foundation models are now evaluated: reasoning capability, coding performance, multimodal understanding, image and video generation, coding ability, context length, agentic tool use, cost per token, deployment flexibility, and enterprise security features. Cohere has carved out a defensible niche in enterprise deployment flexibility and regulated-industry applications, but Canadian consumers or small businesses may continue to largely depend on American or other global model providers. The August 2025 incident in which Anthropic revoked OpenAI's API access to Claude, blocking benchmarking and safety testing, demonstrated that API access can be terminated without notice (Franzen, 2026). Organizations that build critical workflows around foreign model APIs face service disruption risk that no contractual provision fully mitigates.

Societal sovereignty faces substantive risk through model values and political norms embedded in foundation models. AI models are cultural products embedding creator biases (Escott, 2025). Models trained predominantly on American content encode American assumptions about appropriate speech, historical narratives, legal concepts, and social norms. For a bilingual, multicultural country, this matters.

For example, American political debates directly shape how these models respond to Canadian users. In February 2024, Google paused Gemini's image generation model after social media backlash over perceived political bias in its outputs (Axios, 2024). In July 2025, President Trump signed an executive order on "preventing woke AI in the federal government" — the first time the U.S. government explicitly sought to shape the ideological behaviour of AI systems — requiring government-used models to adhere to "unbiased AI principles" (The White House, 2025c). Major foundation model companies responded by emphasizing their commitment to political neutrality. The key insight for Canadian sovereignty is not which political direction these modifications take, but that American political pressures, which are defined by American standards of what constitutes "bias," determine how these models respond to Canadian users. Research suggests that true political neutrality in large language models is neither fully attainable

nor universally desirable; models inevitably reflect the cultural and political contexts of their creators (Hall et al., 2025; ScienceDirect, 2025).

Vulnerability assessment: Moderate. Canada can be resilient through diversification. Cohere provides a genuine Canadian alternative with specific strengths in enterprise deployment. Open-source models can enable deployment on sovereign infrastructure without dependency on foreign model providers. However, heavy consumer and employee usage of American models creates dependence that enterprise procurement policies cannot fully address.

Strategic options: Government procurement preferences for Cohere and Canadian model providers, financial support to prevent foreign acquisition, investment in open-source model deployment capabilities, development of domain-specific Canadian models (legal, bilingual, sector-focused), and enterprise policies addressing shadow AI and uncontrolled model usage.

VII.F Model Inference, Operations, and Orchestration

This layer is the bridge between foundation model capability and application services, encompassing everything required to deploy, run, and manage AI models in production: including exposing them via APIs, scaling inference to meet demand, managing latency and throughput, and orchestrating complex AI workflows.

The layer is comprised of several interconnected capabilities. Model serving and inference optimization address how models are exposed to applications, including load balancing, batching, and hardware acceleration. MLOps (machine learning operations) encompasses the broader lifecycle: workflow orchestration, monitoring for model drift, automated retraining pipelines, experiment tracking, and version control. Vector databases and retrieval-augmented generation infrastructure represent where enterprise knowledge meets models: embedding documents for semantic search, storing and querying those embeddings, and reranking results for relevance. Agent orchestration frameworks enable multi-step reasoning and tool use, while operational fine-tuning techniques allow lightweight customisation without full retraining.

The global landscape is divided into two camps. Hyperscaler-integrated platforms such as AWS SageMaker, Google Vertex AI, and Microsoft Azure ML provide managed, enterprise-grade services covering the full MLOps lifecycle, tightly coupled to their respective cloud ecosystems. Against these, a robust open-source ecosystem offers viable alternatives at every capability level: MLflow for experiment tracking, vLLM for high-throughput inference serving, LangChain and LlamaIndex for agent orchestration, and open-source vector databases (Milvus, Weaviate, Qdrant) for retrieval-augmented generation. No Canadian equivalent to SageMaker or Vertex AI exists, but the open-source alternatives run on any infrastructure, including sovereign cloud.

The key observation for sovereignty is that strong open-source presence exists across all Layer 6 components. Organizations can assemble a complete inference and operations stack without proprietary dependencies.

The Canadian MLOps market reached an estimated US$2.8 billion in 2025 and is projected to reach US$11.4 billion by 2031 at approximately 26% compound annual growth, driven by AI adoption across finance, healthcare, retail, and manufacturing (Mobility Foresights, 2025). Emerging sovereign AI infrastructure (TELUS AI Factory and Bell AI Fabric) provides Canadian-controlled endpoints for the full operations stack. Cohere, while primarily a foundation model company (Layer 5), supplies critical Layer 6 infrastructure through its embedding models for enterprise retrieval, reranking models for search relevance, and deployment tooling for production environments.

Technological sovereignty faces moderate pressure. The open-source ecosystem means organizations can assemble a complete inference and operations stack, from model serving through vector search to agent orchestration, without proprietary dependencies. Kubernetes, MLflow,

Pullquote here if needed for visual interest and entry point, pullquote. Dere prerib volup solupiet parita simi

vLLM, and open-source vector databases all run on sovereign cloud infrastructure. The question is not whether sovereign alternatives exist, but whether Canadian organizations adopt them.

Economic sovereignty warrants attention through a subtler mechanism: hyperscaler MLOps lock-in. When organizations adopt SageMaker or Vertex AI for model serving, experiment tracking, and pipeline orchestration, they add switching costs on top of existing cloud infrastructure dependencies. This compounds the vendor lock-in identified at Layer 4. Each additional hyperscaler service adopted — whether for model monitoring, feature stores, or automated retraining — deepens the dependency and raises the cost of migration. Open-source tooling deployed on sovereign infrastructure avoids this compounding effect.

Vulnerability assessment: Low-moderate. Domestic and open-source options exist across all Layer 6 capabilities. Canadian AI trust and governance companies represent genuine and differentiated capability. The primary risk is not technological lock-in at this layer, but rather the indirect deepening of Layer 4 cloud dependencies through hyperscaler MLOps adoption.

Strategic options: Adoption of open-source MLOps and inference tools on sovereign infrastructure to preserve portability, leverage of emerging Canadian sovereign AI infrastructure, and support for Canadian AI trust and governance companies through federal procurement.

VII.G Applications

The application layer is where AI capabilities meet end users and businesses. Three categories define the landscape: direct AI assistants (ChatGPT, Claude) accessed via web or mobile applications; integrated AI tools (Microsoft Copilot in Office 365, Gemini in Google Workspace) embedded within productivity suites; and AI-enabled platforms built by companies that incorporate AI capabilities into their own products and services.

As documented in Section VII.E, the global consumer AI market is dominated by American providers, with ChatGPT and Google Gemini leading in consumer adoption. Moreover, integrated AI tools create bundled dependency: 70% of Fortune 500 companies have adopted Microsoft Copilot, though mostly in pilot phases with data governance concerns as the primary barrier to enterprise-wide deployment (Jae, 2025). Microsoft 365 with Copilot and Google Workspace with Gemini bundle AI capabilities into productivity infrastructure, creating lock-in that extends beyond AI specifically.

The "shadow AI" problem compounds this dependency. Shadow AI refers to employees using personal AI tools (e.g., personal ChatGPT accounts, consumer-grade applications) for work tasks without IT approval or oversight. According to IBM Canada research from September 2025, 79% of Canadian office workers use AI at work but only 25% rely on enterprise-grade tools; the remainder constitutes shadow AI, creating uncontrolled data flows to foreign providers (Pimentel, 2025).

Canadian companies building with AI represent genuine domestic strength at global scale and a strategic asset for sovereignty. Unlike direct AI tools where Canadians are consumers of foreign services, AI-enabled application companies own their customer relationships and can choose their underlying infrastructure. Clio secured a $900 million Series F at a $3 billion valuation in July 2024, the largest venture investment in Canadian tech history, with over 150,000 legal professionals using its platform (Azevedo, 2024). Shopify serves millions of merchants globally with Shopify Magic and Sidekick AI features integrated throughout its commerce platform (Shopify, 2025; Choudhary, 2025; Shopify, 2026). Wealthsimple has reached CAN$100 billion in assets under administration at a $10 billion valuation and achieved profitability in 2024, with AI-powered research and trading tools launching in early 2026 (Scott, 2025b). Ada Support serves companies like Meta and Shopify at a $1.2 billion valuation with AI-powered customer service automation (Scott, 2021). BenchSci serves 16 of the top 20 pharmaceutical companies and over 50,000 scientists worldwide with its AI drug discovery platform (BenchSci, 2023). OpenText, headquartered in Waterloo, is one of Canada's largest software companies by revenue and has integrated AI capabilities across its information management platform, serving enterprises and governments worldwide (OpenText, 2026).

Jurisdictional sovereignty is threatened through shadow AI creating uncontrolled data flows to foreign providers. Consumer-facing generative AI tools like ChatGPT, Microsoft Copilot, and Google Gemini route interactions directly through U.S.-controlled services. For sensitive enterprise and government information, this creates jurisdictional exposure equivalent to cloud infrastructure but without formal contractual controls. According to an IBM study, shadow AI adds approximately CAN$308,000 per data breach in additional costs (IBM Canada, 2025).

Economic sovereignty faces pressure through market concentration. ChatGPT (OpenAI), Claude (Anthropic), Gemini (Google), and Copilot (Microsoft) are all U.S.-based. Bundling with productivity suites extends lock-in beyond AI into core infrastructure. However, mitigating factors exist: the AI application market remains less concentrated than cloud infrastructure, switching costs for end-user tools are lower than for infrastructure services, and market dynamics in the application space are continually shifting as new entrants compete.

Societal sovereignty is implicated because consumer AI adoption shapes how Canadians interact with information, form preferences, and understand their world. When the dominant interfaces are foreign-owned and optimized for foreign markets, the algorithmic mediation of public discourse remains outside Canadian influence. This connects to Section V's warning about AI-amplified democratic distortion.

Vulnerability assessment: Moderate. Consumer and employee AI (both direct assistants and integrated tools) creates high direct foreign dependency through shadow AI and market concentration. AI-enabled application software shows moderate vulnerability; however, Canadian companies have achieved global scale and can choose sovereign infrastructure for their AI deployments, creating a pathway to managed dependency.

Strategic options: Enterprise procurement policies requiring data governance controls for AI applications, development of Canadian-owned alternatives for government use, support for Canadian AI application companies, and incentives for deployment on sovereign infrastructure. Shadow AI can be addressed through enterprise IT policies, user education, and adoption of enterprise-grade tools with appropriate data governance controls.

VII.H Vulnerability Analysis Summary

The following chart maps the seven AI stack layers against the five dimensions of digital sovereignty, indicating both which dimensions are threatened at each layer and the severity of risk to Canadian AI sovereignty.

Table VII.1: Canadian AI Sovereignty Risk Heat Map

	Jurisdictional	Operational	Technological	Societal	Economic
1. Data & Data Governance	LOW	LOW	LOW	MODERATE	LOW
2. Physical Infrastructure	MODERATE	LOW	LOW	LOW	LOW
3. Computer Hardware	LOW	LOW	CRITICAL	LOW	CRITICAL
4. Cloud Infrastructure	CRITICAL	HIGH	HIGH	LOW	HIGH
5. Foundation Models	LOW	LOW	LOW	MODERATE	MODERATE
6. Model Operations	LOW	LOW	MODERATE	LOW	LOW
7. Applications	MODERATE	LOW	LOW	MODERATE	MODERATE

LOW MODERATE HIGH CRITICAL

Table VII.2: Vulnerability Summary Across the AI Stack

Layer	Vulnerability	Key Sovereignty Threats
1. Data Governance	Low-Moderate	Canadian content underrepresented in training datasets; lacks mechanisms to leverage proprietary data assets for sovereign AI while maintaining privacy protections
2. Physical Infrastructure	Low-Moderate	Network routing through U.S. jurisdiction exposes Canadian traffic to foreign legal exposure; significant foreign ownership of data centre capacity
3. Compute Hardware	Critical	Concentrated supply chain (NVIDIA dominance, TSMC fabrication monopoly); vulnerable to export controls and allocation decisions; no viable domestic alternatives
4. Cloud Infrastructure	Critical	U.S. CLOUD Act jurisdiction despite Canadian data residency; service denial "kill switch" risk; vendor lock-in creates switching costs; concentration risk with hyperscalers
5. Foundation Models	Moderate	Single-vendor concentration with Cohere as the only Canadian option; dependency on foreign providers creates revocation risk; American and/or Chinese political norms shape model values and behaviour; heavy consumer/employee usage creates lock-in beyond enterprise procurement
6. Model Operations	Low-Moderate	Hyperscaler MLOps platforms deepen cloud lock-in
7. Applications	Moderate	"Shadow AI" creates uncontrolled data flows to foreign providers; consumer AI concentration (ChatGPT, Copilot, Gemini); Canadian AI-enabled companies strong, but underlying infrastructure choices matter

This assessment reveals vulnerabilities at every layer of the AI stack, but two stand out as requiring urgent attention. Cloud Infrastructure (Layer 4) represents Canada's most acute vulnerability with realistic domestic solutions. Compute Hardware (Layer 3) presents equally severe vulnerability but cannot be addressed through domestic action alone. Together, these critical middle layers concentrate the threats that flow through the entire stack.

VIII. STRATEGIC OPTIONS FOR CANADA'S AI STACK (INFERENCE)

The previous section assessed Canada's sovereign AI capacity layer by layer, revealing critical vulnerabilities at the compute hardware and cloud infrastructure layers, moderate concerns at the foundation model and application layers, and relative strength at the physical infrastructure layer. This section translates that vulnerability assessment into actionable strategic options. It presents options, not a definitive implementation plan, and acknowledges the trade-offs that policymakers might weigh. The goal is to equip decision-makers with a menu of interventions calibrated to risk, feasibility, and Canada's existing legal and trade constraints.

One additional note: this section addresses model access and deployment (inference), not model creation (training). Foundation model training warrants separate treatment in Section IX.

VIII.A Developing Strategic Options

A PRAGMATIC, RISK-BASED APPROACH

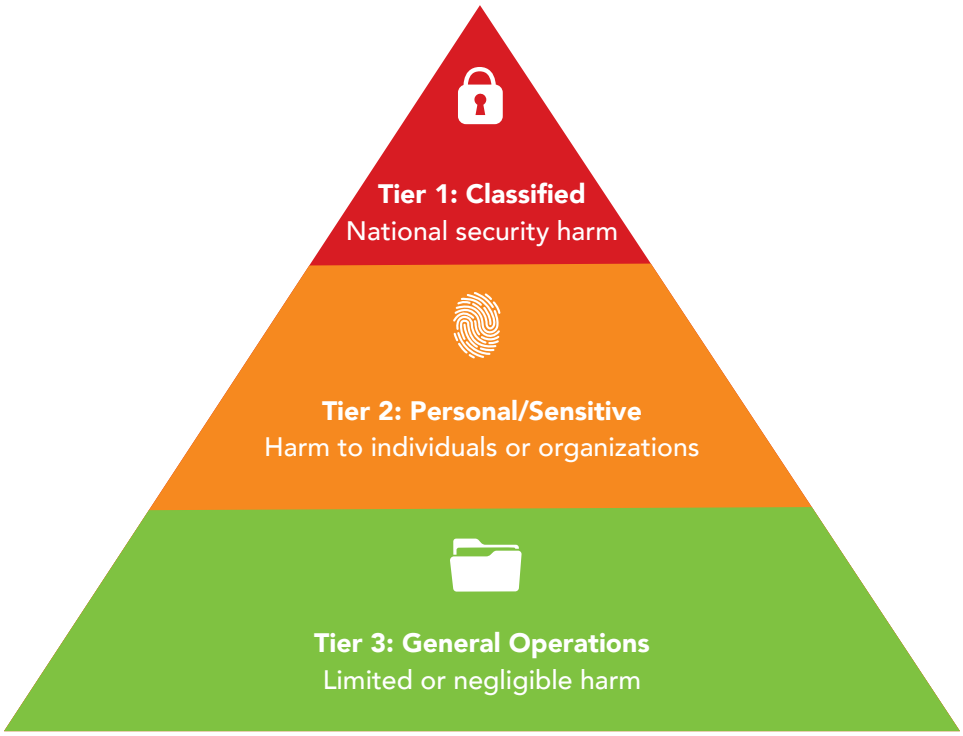
A sovereign AI strategy must be pragmatic and risk-based. Our overarching objective is therefore autonomy and self-sufficiency where possible, but strategic, managed interdependence when necessary. Complete technological independence is neither achievable nor desirable for a middle power. The question is how to structure dependencies to preserve choice and reduce coercion risk, while building genuine domestic capacity wherever feasible.

This pragmatic approach requires acknowledging fundamental trade-offs that run throughout sovereignty policy: cost versus degree of sovereignty (as stated earlier in the report, sovereign infrastructure can often carry a 15–20% premium and the sovereign requirements are often more expensive (Maisto, 2025b)) and innovation versus degree of sovereignty (maximum sovereignty may limit access to cutting-edge capabilities, while maximum access to global solutions may require accepting further dependency and potential coercion risk). Neither extreme serves Canadian interests. The challenge is calibrating the balance appropriately for different contexts.

The layer-by-layer assessment in Section VII revealed that accepting interdependence does not mean accepting vulnerability. Every layer of the AI stack requires strategic attention; what differs is the nature of the intervention. Where Canada can build domestic capacity, as with cloud infrastructure, the strategy emphasizes building and procuring from Canadian-owned or Canadian-controlled providers while also recognizing that global cloud service providers will always have a role. Where domestic solutions are not feasible, as with advanced compute hardware, the strategy emphasizes diversified reliance across allied nations, bilateral agreements that go beyond baseline trade commitments, and leveraging Canadian assets in other domains to secure technology access when necessary.

A KEY TOOL: THE DATA SENSITIVITY SPECTRUM

Navigating the trade-offs between cost, security, and innovation becomes easier when sovereignty requirements are calibrated to how sensitive and critical the operations are. Not all government data requires the same protection; not all AI workloads carry the same consequences if compromised. Appendix A details a tiered framework for calibrating sovereignty requirements to data sensitivity and institutional context (e.g., federal, provincial, municipal, private sector).



This tiered approach ensures that sovereignty measures are proportional to risk. Classified workloads demand the highest standards. General business or routine government operations should require a more nuanced trade-off between sovereignty goals and market conditions. The spectrum helps policymakers avoid two failure modes: under-protection of genuinely critical systems, and over-prescription that imposes unnecessary costs on lower-risk applications.

VIII.B Critical Priority: Sovereign Cloud Infrastructure

Cloud infrastructure (Layer 4) is the critical strategic intervention point. Section VII identified it as Canada's most acute controllable vulnerability, with high exposure across four sovereignty dimensions but realistic domestic alternatives. Unlike AI compute hardware, where foreign supplier reliance cannot be resolved through Canadian action alone, sovereign cloud infrastructure is achievable. Canada has domestic providers, government procurement leverage, and viable legal pathways.

Three models present a spectrum of approaches to strengthening Canada's AI sovereignty in the context of cloud infrastructure, from maximum sovereign ownership to maximum flexibility. The appropriate choice depends on both the sensitivity of the workload and the institutional context governing it.

Self-Hosted Infrastructure: The option with the highest sovereignty level is government-owned and operated infrastructure with no reliance on foreign technology licensing. Self-hosted environments are typically air-gapped from public networks, operated exclusively by cleared government personnel, and built on fully audited technology. This approach eliminates foreign jurisdictional exposure entirely with no foreign legal orders, no service denial risk, and no supply chain dependencies on commercial providers. The disadvantages, however, are substantial: the highest cost, the most limited access to innovation, and significant operational burden. Self-hosting is neither efficient nor scalable for most workloads, but it may be necessary for the most sensitive classified operations, particularly for “Canadian Eyes Only” data and workloads.

Juridical Cloud Sovereignty (French/German Model): This approach requires private sector Canadian-owned and operated infrastructure, creating a legal "air gap" from foreign jurisdiction. The defining feature is that a Canadian entity, not a foreign parent company,

has legal and operational control over the infrastructure, meaning foreign legal orders cannot compel access to Canadian data even if foreign technology underlies the platform.

This model encompasses Canadian ownership and operation by a Canadian legal entity, with all personnel holding Government of Canada security clearances, and data residency in Canada with no exposure to foreign legal orders. Licensed operator arrangements, where a Canadian entity operates foreign technology under Canadian control, also qualify. France's Bleu cloud, a partnership between Orange, Capgemini, and Microsoft where the French joint venture operates Microsoft technology under French sovereignty, exemplifies this approach (Capgemini & Orange, 2025). Similarly, Delos, a subsidiary of the German company SAP which is subject to German control, provides Microsoft Azure and Office 365 capabilities, representing another implementation (European Cloud, 2025). In both cases, the foreign technology provider supplies the software platform, but the sovereign entity maintains operational control, holds encryption keys, and employs the personnel who administer the system.

The advantages are clear: maximum control and complete jurisdictional separation from foreign legal jurisdiction. The Canadian operator, not a foreign parent company, possesses and controls Canadian data. The disadvantages are equally clear, including the higher cost as sovereign cloud typically carries a 15–20% premium (Maisto, 2025a). There could also be a potential technology lag as licensed technology must be adapted and validated for sovereign operation, and limited provider options initially as the market develops.

Contractual Cloud Sovereignty (Australian Model): This approach achieves sovereignty through outcome-based contractual controls rather than ownership requirements. Any provider, including foreign hyperscalers, can qualify if it meets stringent requirements: all sensitive data encrypted with keys held exclusively by Canadian entities, all administrative access limited to personnel with Government of Canada security clearances, contractual "step-in" rights if provider ownership changes, and full audit rights for Canadian security assessors.

Under this model, the focus shifts from who owns the infrastructure to how the infrastructure is operated. Foreign hyperscalers could qualify for sensitive government workloads if they invest in meeting the specified outcomes. Australia's Hosting Certification Framework can provide a template. Through it, providers achieve certification by demonstrating that they meet defined security and sovereignty outcomes, regardless of corporate nationality (Australian Government, 2021).

The advantages include maximizing access to hyperscaler innovation and cost efficiency, avoiding the technology lag associated with licensed operator arrangements, and trade-agreement defensibility because requirements are origin-neutral (specifying outcomes, not nationality). The disadvantage is that contractual controls cannot fully address extraterritorial legal reach. A U.S. legal order served on a U.S. company creates compliance obligations that contracts with Canadian customers cannot override (Appleton, 2025). However, Canadian-held encryption keys through hardware security modules operated by cleared personnel can create a practical barrier: if a Canadian customer holds the only decryption keys through hardware security modules, a provider served with a U.S. court order would likely be technically unable to comply. This creates a meaningful practical barrier to foreign legal access even without juridical separation (but does not eliminate the underlying legal obligation).

DETERMINING MINIMUM SOVEREIGNTY LEVELS

To help further ground these options, Table VIII.1 presents one feasible set of minimum sovereignty levels to each combination of data sensitivity tier (from Appendix A.II) and institutional context. The assignment reflects both the risk profile of the workloads and the policy levers available in each institutional context.

Table VIII.1: Minimum Sovereignty by Tier and Institutional Context

Tier	Federal Government	Provincial/Municipal	Private Sector	
			Federally Regulated Industries and CCSPA sectors	General Private Sector
Tier 1 (Classified)	Self-hosted or Juridical	N/A	N/A	N/A
Tier 2 (Personal/ Sensitive)	Juridical	Juridical	Contractual	Contractual
Tier 3 (General Operations)	Contractual*	Contractual*	Marketplace Determined	Marketplace Determined

**In practice, there is potential to pool government demand across tiers and/or levels of government. See discussion on pooling below.*

These minimums reflect a key trade-off. Self-hosted infrastructure provides the highest levels of sovereignty and is a viable option for workloads that demand it. However, its cost, performance, and innovation constraints may be disproportionate where juridical sovereignty provides adequate jurisdictional separation. Decision-makers should seriously consider whether full self-hosting is necessary for a given workload or whether juridical sovereignty, with its Canadian ownership, cleared personnel, and legal separation from foreign jurisdiction, delivers sufficient protection. The table identifies the lowest standard sufficient for each combination of sensitivity and institutional context; organizations are always free to exceed the minimum.

Tier 1 (Classified). Classified workloads face the highest-consequence threats: espionage, sabotage, and foreign intelligence targeting. The legal exposure and service denial risks documented in Section VII make anything less than sovereign ownership insufficient.

Within Tier 1, self-hosted infrastructure is likely most appropriate for "Canadian Eyes Only" workloads, where eliminating most foreign technology dependencies provides the highest assurance. Juridical sovereignty is appropriate for systems that must be interoperable with allied intelligence partnerships, where Canada's position as a net consumer of foreign intelligence requires maintaining technical compatibility. Tier 1 workloads are almost exclusively held at the federal level. The rare exceptions — shared classified data held by provinces and classified contracts held by defence and intelligence contractors — would likely follow federal Tier 1 requirements.

Tier 2 (Personal/Sensitive). Sensitive personal and organizational data, such as health records, financial information, and personnel files, pose a significant threat if breached, but not at the national security level of Tier 1. For federal government data, juridical cloud sovereignty is the recommended minimum

given available procurement levers and the sensitivity of the workloads involved. For provinces and municipalities, which hold the bulk of Tier 2 data (health, education, social services, vital statistics), juridical sovereignty is also the recommended minimum.

The recommended minimum for private sector organizations dealing with Tier 2 data is contractual rather than juridical sovereignty. This reflects practical considerations. Public expectations of government data stewardship are higher: citizens entrust their most sensitive information to the government on the understanding that it will be protected to the highest feasible standard. The private sector is also far more fragmented with thousands of organizations of varying size, sophistication, and risk profile, making a uniform juridical sovereignty mandate impractical. Organizations that deem a higher standard necessary remain free to adopt juridical sovereignty voluntarily.

A contractual minimum is nonetheless important for Tier 2 data held by the private sector. Sensitive personal information including health records, financial data, and biometric identifiers like fingerprints, warrant baseline protections regardless of whether the custodian is a government department or a private firm; the harm from a breach does not diminish because of who holds the data. For CCSPA-regulated industries, the federal government already has regulatory levers through OSFI, the CRTC, and the Canadian Energy Regulator. For the broader private sector, a contractual sovereignty floor could be established through sector-by-sector regulation or through legislation modelled on the Critical Cyber Systems Protection Act. In both cases, consultations with the private sector should precede any regulatory or legislative action.

Tier 3 (General Operating). General operating data and routine applications prioritize cost efficiency and access to innovation. Most organizations at this tier will use commercially available cloud services. For government workloads (federal, provincial, and municipal), the recommended minimum is contractual sovereignty, ensuring a baseline of Canadian-held encryption keys and cleared personnel even for lower-sensitivity workloads. For federally regulated industries, adoption should be determined by the marketplace, although specific nuances could be determined on a sector-by-sector basis by regulators and industry associations. For the general private sector, adoption should be marketplace-determined; no minimum is mandated, though many enterprises may voluntarily choose contractual or juridical options for competitive, reputational, or risk management reasons.

Pooling for Critical Mass & Market Development

The minimum sovereignty levels in Table VIII.1 represent a risk-based floor. Practical considerations argue for exceeding the minimum in certain cases, particularly through pooling demand across tiers and institutional boundaries.

The pooling model depends on a critical design choice: juridical sovereignty infrastructure should be operated by Canadian entities that meet juridical sovereignty criteria but are arms-length from the federal government. This is not a model where other levels of government entrust their data to Ottawa for safekeeping. Rather, an independent Canadian sovereign compute provider, or consortium of providers, would satisfy juridical sovereignty requirements while serving clients across federal, provincial, and municipal governments. The Bleu model in France, where a sovereign joint venture operates independently of the French government while serving government clients, provides a useful analogy (Capgemini & Orange, 2025).

For the federal government, segmenting Tier 2 and Tier 3 data onto separate infrastructure may not be worthwhile. The administrative complexity and fragmented demand could possibly outweigh any cost savings from using a lower standard for Tier 3 workloads. Pooling both tiers onto juridical sovereignty infrastructure could simplify procurement, build critical mass for domestic providers, and reduce per-unit costs by narrowing the sovereignty premium. The same logic could apply to provinces and municipalities. Rather than maintaining separate arrangements for different sensitivity tiers, consolidating onto sovereign compute infrastructure is a potential option for achieving better economies of scale.

In addition, beyond pooling within a single level of government, there could be significant benefits in pooling across levels of government. If the federal government, provinces, and municipalities access the same arms-length sovereign compute

providers, the combined demand strengthens the anchor buyer effect and makes sovereign alternatives commercially viable at a scale that no single jurisdiction could sustain alone.

A sovereignty strategy designed tier-by-tier and jurisdiction-by-jurisdiction risks fragmenting demand across too many arrangements. Consolidating workloads onto shared sovereign infrastructure, operated by arms-length providers that satisfy juridical sovereignty criteria, creates the volume needed to sustain competitive domestic providers.

Key Policy Mechanisms

Four policy mechanisms enable sovereign cloud development, each requiring deliberate attention and investment.

1. Procurement reform through updating the Government of Canada Cloud Framework with explicit sovereignty tiers and evaluation criteria. Current procurement vehicles do not adequately distinguish between sovereignty levels or provide clear pathways for Canadian providers to demonstrate compliance. Streamlined certification processes should enable Canadian providers to demonstrate compliance efficiently without prohibitive administrative burden. See Section VIII.F for additional discussion of procurement reform as a cross-cutting enabler.

2. Application of the National Security Exception where appropriate. The SSC Sovereign Cloud Request for Information demonstrates that Canada has both recognized the vulnerability and identified the legal pathway; the policy challenge is building the domestic capacity to deliver on sovereign cloud commitments.

3. Certification frameworks enabling Canadian alternatives to demonstrate that they meet security requirements. Standardized assessment criteria should allow any provider to understand what is required, invest in meeting requirements, and achieve certification through a predictable process. Certification should address both initial qualification and ongoing compliance validation.

4. Direct support for domestic sovereign compute through procurement opportunities, financing, and potential equity contributions for domestic sovereign compute consortiums. Policy should prioritize government funding for domestic sovereign compute capacity over general data centre build-out. Government as the anchor buyer establishes market critical mass; direct support can accelerate capability development where market signals alone are insufficient.

VIII.C Critical Dependency: Compute Hardware

Compute hardware (Layer 3) presents a vulnerability as severe as cloud infrastructure, but requires a fundamentally different strategy because comprehensive domestic solutions are not feasible. The threat is also less immediate: a supply chain disruption limits the ability to build new capacity but does not compromise systems that are already deployed and operational.

Since domestic production is impossible, the goal is strategic interdependence: diversifying dependencies to preserve choice and reduce vulnerability to foreign leverage. This strategy has four elements.

1. Supply chain diversification across allied nations. Canada should avoid concentration of dependency on any single country by cultivating relationships with multiple suppliers across allied nations and monitoring the landscape for alternatives as they emerge. Google's TPUs, AMD's accelerators, and emerging RISC-V architectures represent diversification opportunities even if none currently match NVIDIA's ecosystem (Dey, 2025).

2. Bilateral agreements beyond baseline trade commitments. Canada should negotiate supply assurance agreements with key allies; particularly the United States given current dependencies, but also directly with Korea and Taiwan, where critical fabrication capacity is concentrated. Agreements should link Canadian strengths in other domains (energy exports, critical minerals) to technology access when necessary. Beyond securing priority allocation commitments for Canadian government and critical infrastructure needs, bilateral agreements should aim to establish frameworks that reduce the risk of Canada being subjected to export controls as a coercion vector. The United States has demonstrated willingness to use export controls as instruments of foreign policy (Gertz & Krueger, 2025); agreements that clarify

Canadian access and establish consultation mechanisms before restrictions are imposed would reduce vulnerability to unilateral action.

3. Strategic participation in multilateral semiconductor initiatives. Canada should engage with allied coordination efforts like the Chip 4 Alliance dialogue, align Canadian research funding with allied priorities, and explore partnership opportunities in allied semiconductor programs. Canada's exclusion from the Chip 4 Alliance represents a diplomatic gap worth addressing (Naik, 2025).

4. Contingency stockpiling. Strategic reserves of GPUs and critical components for Tier 1 and Tier 2 systems provide resilience against supply disruptions. Coordination with allied stockpiling efforts can reduce costs and improve collective security.

Canada cannot achieve hardware self-sufficiency, but can contribute meaningfully to the global supply chain in specialized areas. Some emerging Canadian companies occupy specialized segments. Supporting these companies and positioning Canadian research in areas valued by allies strengthens Canada's position in negotiations over supply assurance.

VIII.D Model Access and Deployment Strategy

Layers 5, 6, and 7 — foundation models, model operations, and applications — address how Canada accesses and deploys AI capability. This is distinct from Section IX's focus on model creation. The strategy at these layers must balance three objectives: ensuring access to frontier capabilities for Canadian competitiveness, reducing vulnerability to foreign leverage, and building domestic capacity where feasible.

Foundation Model Access (Layer 5)

Three approaches preserve optionality at this layer, each addressing different aspects of the vulnerability profile identified in Section VII.

1. Procurement preferences for Canadian providers. Government procurement should prefer Canadian models and AI solutions where security requirements permit. For Tier 2 workloads, this could go as far as explicit mandates for Canadian foundation models in procurement requirements. The rationale extends beyond nationalism: when models can be deployed on sovereign infrastructure without data leaving Canadian jurisdiction, the CLOUD Act exposure that defines Layer 4 vulnerability does not propagate to Layer 5. Procurement preferences also create a virtuous cycle: Canadian government adoption generates revenue that funds continued development, which improves model capability, which strengthens the case for further adoption.

2. Open-source models as a strategic hedge. Sovereign infrastructure should support deployment of open-weight models such as Llama and Mistral to preserve optionality and reduce single-provider dependency. Open-weight models can be deployed entirely on Canadian-controlled infrastructure, eliminating foreign model dependency and revocation risk. Small open-source models represent a particularly valuable strategic option for many government and enterprise use cases. For example, Mistral's open-source offerings include smaller, more efficient models that can run on modest infrastructure while delivering capable performance for routine tasks (e.g. document processing, summarization, translation) and similar workloads that constitute the bulk of practical AI deployment (Laurent, 2026). These smaller models enable sovereign AI deployment without requiring frontier-scale compute, making sovereignty achievable for a broader range of applications and organizations.

However, it is important to acknowledge a key limitation: open-source models currently lag closed frontier models in reasoning, coding, and agentic capabilities. This makes open source a hedge rather than a primary strategy; a fallback that ensures Canada retains capability even if commercial relationships deteriorate. The strategic implication is that sovereign infrastructure should be architected to support multiple model sources, with workloads that can migrate between proprietary and open-source models as capability and circumstances evolve.

3. Diversification across model sources. Critical government workloads should avoid single-provider dependency and architect for model portability wherever possible. This means standardized interfaces that abstract model-specific APIs, containerized deployment

patterns that enable workload migration, and retention of model weights rather than pure foreign model dependency. The goal is ensuring that no single provider, domestic or foreign, can exercise unilateral leverage through service denial or unfavourable terms. This diversification principle applies even to Cohere: while procurement preferences support the Canadian provider, architectures should not create lock-in that would leave the government vulnerable if Cohere's circumstances change.

Model Operations (Layer 6)

Section VII assessed this layer as low-moderate vulnerability because strong open-source alternatives exist. The strategic approach emphasizes portability and domestic capability.

The federal government should adopt and promote open-source MLOps and inference tooling on sovereign infrastructure wherever possible. These tools provide enterprise-grade capabilities without proprietary lock-in to hyperscaler platforms. The key insight is that sovereignty at this layer derives primarily from Layer 4 infrastructure choices: if you control the cloud, you can control the operations stack. Hyperscaler MLOps tooling deepens cloud lock-in by creating additional switching costs beyond the infrastructure itself.

There are many Canadian AI governance and trust companies. Examples like Private AI for data protection, Armilla for AI assurance, TrojAI for adversarial protection, and Integrate.ai for federated learning represent genuine domestic capability in an increasingly important segment. Federal procurement should actively support these companies, both to develop domestic capacity and to ensure that AI governance requirements can be met with Canadian solutions.

Applications (Layer 7)

The application layer presents complex concerns requiring different policy responses.

Shadow AI (employees using personal AI tools without IT approval) creates uncontrolled data flows to foreign providers. Section VII documented that 79% of Canadian office workers use AI at work but only 25% rely on enterprise-grade tools (Pimentel, 2025). The remainder routes sensitive information through personal accounts on foreign platforms, creating jurisdictional exposure that enterprise procurement cannot control. Addressing shadow AI requires enterprise policies mandating approved AI tools with appropriate data governance controls, combined with user education about the risks of shadow AI and the adoption of enterprise-grade alternatives that meet user needs. The goal is not to prohibit AI usage but to channel it through controlled pathways.

Canadian AI-enabled companies and tech leaders including Shopify, Clio, Wealthsimple, Ada Support, BenchSci, Open Text, and many others represent genuine domestic strength at global scale. These companies own their customer relationships and can choose their underlying infrastructure. Supporting their deployment on sovereign infrastructure creates anchor demand while demonstrating commercial viability. Incentives for sovereign infrastructure adoption among Canadian AI companies serves a dual purpose: it reduces the companies' own exposure to foreign leverage while creating domestic market demand that supports sovereign infrastructure development.

VIII.E Supporting Infrastructure

Physical Infrastructure (Layer 2)

Section VII assessed this layer as relatively strong given Canada's green energy advantage and growing domestic data centre capacity. Strategic options focus on reducing remaining vulnerabilities while leveraging existing strengths.

Network routing optimization addresses the jurisdictional exposure created when Canadian domestic traffic transits through the United States. As mentioned in VII.B, more than 25% of Canadian domestic internet traffic routes through U.S. exchange points, exposing it to surveillance authorities (IXmaps, 2026). For sensitive government and critical infrastructure communications, end-to-end encryption should be mandated for any traffic traversing foreign infrastructure, ensuring that even if routing cannot be controlled, content remains protected.

Energy leverage represents Canada's most significant physical infrastructure advantage. Roughly 80% of Canadian electricity is non-

emitting and 73% is hydro or other renewables, positioning the country as one of the world's most attractive locations for AI infrastructure investment (Government of Canada, 2025). This advantage should be leveraged strategically: regulatory incentives can condition access to Canadian clean energy on sovereignty-enhancing commitments, including domestic data residency, local hiring, and participation in Canadian supply chains. The goal is not to restrict investment but to ensure that Canada's energy advantage translates into sovereignty benefits rather than merely hosting foreign-controlled capacity.

Canadian-owned data centre capacity provides alternatives to hyperscaler facilities. eStruxture, QScale, and telecommunications providers operating data centres offer domestic options for organizations that require Canadian ownership. Policy should support expansion of this capacity while ensuring it can compete effectively with hyperscaler offerings on cost and capability.

Data and Data Governance (Layer 1)

Canadian datasets (e.g., health data through CIHI, financial data through OSFI, administrative data through Statistics Canada and CRA, legal data through CanLII) represent strategic assets requiring preservation and sovereign control. These datasets could provide significant value for AI development, but mechanisms to enable responsible access while maintaining privacy and sovereignty are underdeveloped. Canada's data laws and frameworks are long overdue for modernization, given it has been decades since major modernization and in many cases the laws originate from before modern technologies.

Government privacy laws and data governance frameworks should be updated to enable AI development while protecting Canadians' data. This means clear classification policies determining which government datasets can be used for AI training and fine-tuning, under what conditions, and with what controls. The goal is unlocking the value of Canadian data assets for Canadian AI development rather than allowing them to remain inaccessible while foreign models trained on other datasets dominate Canadian markets.

Addressing Canadian content underrepresentation in global training datasets preserves societal sovereignty. Quebec French accounts for less than 5% of French-language recordings in Common Voice (Serrand et al., 2025). When AI models systematically exclude Canadian content, they embed foreign assumptions about language, culture, and norms. Incentives for Canadian content creation and coordination with training dataset curators can address these representational gaps.

VIII.F Cross-Cutting Enablers

Four policy mechanisms span multiple layers of the AI stack. Each requires sustained attention and investment to enable the strategic options outlined above.

1. Procurement Reform

Across the AI stack, current procurement vehicles do not adequately distinguish between sovereignty levels. Reform should establish criteria that account for sovereignty benefits (Canadian ownership, personnel with clearances, jurisdictional separation) alongside traditional factors of cost and capability.

More specifically, procurement reform touches three layers of the AI stack. At the cloud infrastructure layer (Layer 4), the Government of Canada Cloud Framework Agreements require updating with explicit sovereignty tier requirements and standardized assessment criteria. At the foundation model layer (Layer 5), procurement should create pathways for Canadian model providers to compete for government contracts on terms that reflect the sovereignty value of domestically developed models. At the application layer (Layer 7), Canadian AI and software companies have encountered persistent difficulty navigating government procurement. Recent research documents demonstrate that federal IT procurement "betrays accepted best practice" on contract values, supplier diversity, and intellectual property management. These are patterns argued to be "a historically overlooked but crucial driver of [government's] failing digital reform efforts" (Boots et al., 2024).

The Buy Canadian Policy announced in December 2025 provides a starting point, but ICT-specific guidance is needed to translate general preferences into actionable procurement requirements for cloud and AI services. In addition to potentially using the National Security Exception, one option may be to categorize

AI infrastructure and critical data centres as dual-use infrastructure under the Policy on Prioritizing Canadian Materials in Federal Procurement (PSPC, 2025d). Reform should also entail streamlined certification processes so that Canadian providers can efficiently demonstrate they meet sovereignty requirements without prohibitive compliance burdens.

Ideally, procurement requirements would map directly onto the three-tier Data Sensitivity Spectrum outlined in Appendix A. This alignment ensures that procurement criteria are proportionate to risk rather than applying a single standard across all workloads.

2. Workforce and Security Clearance Capacity

Sovereign cloud and AI infrastructure requires skilled, qualified personnel with appropriate security clearances at every tier. The federal government should assess clearance processing capacity against projected sovereign cloud staffing needs and address bottlenecks before they constrain implementation. Section VII documented a 30% vacancy rate in federal digital roles (Treasury Board of Canada Secretariat, 2025c). Sovereign AI initiatives will likely compete for the same limited talent pool within the government.

The challenge is threefold. First, the AI and cybersecurity talent pipeline requires coordinated development across educational institutions, immigration pathways, and professional development programs. Second, clearance processing must keep pace with demand: if sovereign infrastructure requires cleared personnel to operate, delays in the security screening process become delays in sovereign capacity. Third, Canada's security classification framework requires modernization; as noted in Appendix A, the current framework maps imperfectly to sovereignty tiers, and allied models from the United Kingdom and Australia offer instructive approaches to calibrating vetting intensity to classification level. Without adequate cleared personnel, sovereign infrastructure cannot operate at the scale required for government workloads, and staffing constraints will limit implementation regardless of policy intent.

3. Data Portability Requirements

The federal government should introduce data portability requirements for cloud providers to ensure that switching providers remains practically feasible, not merely theoretically possible. Section VII documented switching costs reaching 18 months and $8.5 million in some cases. These barriers make the technological sovereignty test ("can you migrate if needed?") fail in practice. Interoperability standards enabling switching without functionality loss should be procurement requirements for any cloud contract. Ukraine's deliberate design of cloud-agnostic systems can serve as a model. Canadian procurement should mandate architectures that preserve migration options, with demonstrated portability as a contract requirement rather than an aspiration.

4. Sovereignty Audit Capabilities

Canada must develop domestic capacity to assess and validate sovereignty claims by cloud and AI providers. Standardized assessment frameworks should reference the Digital Government Standards Institute's work and Canadian AI Safety Institute criteria. Certification processes for sovereign cloud providers should enable efficient validation while maintaining rigorous standards. Ongoing sovereignty validation, not just initial certification, ensures that providers continue to meet requirements as circumstances evolve. Building domestic capacity for these assessments, rather than relying entirely on third-party certifiers, ensures that sovereignty evaluation serves Canadian policy objectives rather than industry preferences.

VIII.G CUSMA Review Preparation

The scheduled CUSMA review in July 2026 provides both opportunity and risk for Canada's digital sovereignty position (Bitar et al., 2025). Section VI documented aggressive U.S. positioning (CCIA, 2025b; Malone, 2025). The options developed in this section provide tools for navigating these negotiations in the context of AI and digital sovereignty.

Play defence on digital sovereignty. Canada should prepare for U.S. pressure on data localization and procurement preferences by developing comprehensive legal defences grounded in the National Security Exception (Global Affairs Canada, 2022, art. 32.2) and government procurement carve-outs (Global

Affairs Canada, 2020, art. 19.2(3)) documented in Section VI. Building a coalition with like-minded partners such as Australia, the U.K., Mexico, the European Union, and other middle powers facing similar pressures can strengthen Canada's negotiating position. The EU's own struggles with American digital dominance, documented in Section III, suggest natural alignment even if Europe's regulatory approach differs from Canada's. Coordinated messaging about legitimate sovereignty concerns, distinguished from protectionism, makes individual country positions harder to dismiss as outliers.

Resist concessions that weaken sovereignty. Protecting national security-based exemptions should be a red line. The National Security Exception provides the legal foundation for measures like the SSC Sovereign Cloud procurement (Shared Services Canada, 2025b); constraining this exception would eliminate the primary pathway for sovereignty-enhancing policy (Global Affairs Canada, 2022, art. 32.2; Government of Canada, 2025). Similarly, preserving government procurement flexibility under Article 19.2.3 is essential: this carve-out enables the outcome-based requirements that make sovereign cloud achievable without violating trade disciplines (Global Affairs Canada, 2020, art. 19.2(3)). Canada should maintain space for outcome-based sovereignty requirements that specify what providers must do (security standards, data handling, audit access) rather than where they must be headquartered. These requirements are trade-defensible precisely because they are origin-neutral (Global Affairs Canada, 2020, art. 19.4).

Avoid trading digital rights for unrelated concessions. Digital sovereignty measures should not become bargaining chips for economic concessions in unrelated sectors. Section VI documented the Digital Services Tax capitulation, where Canada rescinded the tax to resume trade negotiations (Department of Finance Canada, 2025), illustrating the pressure to sacrifice digital policy for broader trade relationship management. The strategic importance of AI sovereignty exceeds the economic value of any near-term tariff relief. Negotiators should be explicitly instructed that digital sovereignty provisions are not tradeable for concessions in automotive, agriculture, or other sectors.

Maintain trade compliance throughout. The strategic options developed in this section are designed to be trade-defensible. Outcome-based requirements that apply equally to domestic and foreign providers do not constitute origin discrimination under CUSMA Chapter 19 (Global Affairs Canada, 2020). The National Security Exception is self-judging and does not require emergency circumstances (Global Affairs Canada, 2022, art. 32.2; Government of Canada, 2025). Government procurement is explicitly carved out from digital trade disciplines (Global Affairs Canada, 2020, art. 19.2(3)). Maintaining compliance demonstrates that Canadian measures serve legitimate sovereignty objectives rather than disguised protectionism, strengthening Canada's legal position if challenged and its credibility in negotiations.

Consider offensive opportunities. The CUSMA review framework includes the possibility that parties may not agree to extend the agreement, triggering termination in 2036 (Bitar et al., 2025) or even a potentially earlier withdrawal. While continuity of trade rules benefits Canada in many respects, the digital trade provisions of Chapter 19 disproportionately benefit American technology companies operating at scale in Canadian markets (Jansa, 2025; CCIA, 2025b). CUSMA's potential modification or elimination would significantly reduce the trade rights of U.S. hyperscalers in Canada, creating policy space for sovereignty measures that current rules constrain. This is not an argument for withdrawing from CUSMA, but recognition that digital sovereignty provisions can be useful leverage in negotiations. Canada should approach the review understanding that the status quo is not necessarily in Canadian interests and that willingness to accept disruption strengthens our negotiating position.

IX. FOUNDATION MODEL DEVELOPMENT AND FUNDAMENTAL AI RESEARCH

The previous sections addressed how Canada can access and deploy foundation models through sovereign inference infrastructure. This section confronts another fundamental question: who creates the models Canada relies upon? Even with robust sovereign cloud capacity and secure deployment options, Canada remains strategically vulnerable if every foundation model it uses originates from actors who could, at their discretion, revoke access, restrict functionality, or embed values antithetical to Canadian interests.

IX.A The Scale Challenge for Model Development

Section VII.E established that Canada could achieve meaningful sovereignty over model inference and deployment through a combined use of Cohere and open-source models and sovereign cloud infrastructure. Model creation presents a fundamentally different challenge: extraordinarily expensive model training costs make it extremely difficult to stay competitive with frontier labs in producing the latest state-of-the-art models.

Training costs have grown at approximately 2.4 times per year since 2016 (Cottier et al., 2024). OpenAI reportedly spent $5 billion on research and development compute in 2024 alone (You, 2025). According to a joint report from leading researchers and institutions (Abecassis et al., 2025), training runs could cost $6.6 billion or more by 2028, with the compute clusters required to support them reaching $20.6 billion. To rise to this challenge, Microsoft and OpenAI's Project Stargate envisions a $100 billion supercomputer initiative (OpenAI, 2025b), while Meta's 2025 capital expenditure ranges between $70 and $72 billion (Meta Platforms, Inc., 2025).

This escalating cost drives concentration in model creation, reducing the number of credible companies that can compete at the frontier to a handful, predominantly in the United States and China. For Canada, this concentration creates a direct sovereignty problem at the model creation layer. Model creators control access, pricing, and terms of use. Even ostensibly open API access can be revoked or conditioned. As one Canadian high performance compute expert noted, "Canadian data—hospitals, banks, military—is trained and stored on infrastructure controlled by U.S. technology giants, which ultimately answer to the U.S. government" (Grant, 2025).

Some argue that algorithmic efficiency gains will democratize frontier capability. Anthropic claims frontier performance with a fraction of OpenAI's compute (Monica & Pratama, 2026), and DeepSeek achieved potentially up to an 18-fold training cost reduction (Noffsinger et al., 2025). However, efficiency gains tend to extend the frontier rather than alter the underlying cost structure for competition at the bleeding edge. The Jevons paradox — the observation that increased efficiency in resource use often leads to increased total consumption rather than conservation — suggests reduced costs may simply stimulate greater demand for ever-larger training runs, pushing the frontier further out of reach (Luccioni et al., 2025).

Middle powers are faced with what appears to be a binary choice: dependency on U.S. or Chinese AI systems with attendant risks of data exposure, service denial, and embedded values, or technological weakness that undermines economic productivity, scientific discovery, and military capability. Neither option is acceptable.

Canada should reject this false choice through a dual approach. The first prong supports Cohere as a domestic foundation model capable of meaningful scale. The second prong pursues multinational partnership to achieve frontier capability collectively. Alone, neither prong is sufficient: a single national champion still leaves Canada vulnerable, while a multinational partnership without domestic capability leaves Canada a passive beneficiary rather than an active contributor. Together, they provide strategic redundancy.

IX.B Supporting Canada's Domestic Foundation Model Capacity

Canada possesses something few middle powers can claim: a significant foundation model company headquartered on domestic soil. Section VII.E examined Cohere's operational capabilities in detail: its enterprise focus, deployment flexibility, government partnerships, and air-gapped infrastructure options. Here, we consider its strategic significance as a foundation model creator for Canadian sovereignty.

Cohere's value lies not in competing head-to-head with OpenAI or Anthropic for frontier dominance, but in providing Canada with at least one foundation model provider subject to Canadian law, Canadian priorities, and Canadian control. Few comparable companies exist outside the United States and China. France has Mistral and Germany has Aleph Alpha, though it has pivoted from model development to an "AI operating system" strategy (Lomas, 2024). Cohere represents a rare asset that ensures Canada is not entirely dependent on foreign model providers for AI capability. Importantly, Cohere's leadership has publicly committed to maintaining Canadian headquarters and framed this commitment in explicitly sovereignty-oriented terms (Galea & Silcoff, 2025).

Supporting domestic foundation model capacity means ensuring Cohere has the resources to remain competitive in an industry where training costs are rising exponentially. This requires government support that goes beyond adopting Cohere for inference purposes to include research and development funding, talent retention incentives, and integration with sovereign compute infrastructure. Canada should consider treating Cohere as a national champion in the tradition of France's support for Mistral, recognizing its unique strategic importance while remaining attentive to the risks that national champion strategies can carry. In addition, greater adoption of Cohere's models by Canadian businesses and government creates a virtuous circle: increased revenue enables investment in better models, which attracts more customers and talent, generating further revenue for reinvestment.

Finally, sovereignty is not exclusively a Canadian concern. Many countries seeking to hedge against U.S. and Chinese AI dominance may prefer a foundation model provider headquartered in a trusted middle power with strong rule of law. To the extent that Cohere can serve this international demand, Canadian sovereignty interests and commercial viability reinforce each other.

IX.C A Multinational Frontier AI Partnership

A multinational partnership offers the only realistic path for countries outside the United States and China to achieve genuine frontier model creation capability. Canada is well-positioned to help lead such an effort.

In November 2025, a consortium including Yoshua Bengio (of the Mila – Quebec AI Institute), the Oxford Martin School, and institutions across France, Germany, and Sweden published "A Blueprint for Multinational Advanced AI Development" (Abecassis et al., 2025). Its core thesis holds that "mid-sized economies likely face insurmountable barriers to independent frontier AI development." However, collective action can achieve what individual countries cannot. The report proposes a semi-distributed structure with equitable cost-sharing, commitment to responsible development, and agile governance.

The "AI bridge powers" concept captures this strategy. Eighty-seven of the 100 most-cited AI researchers have ties outside the United States and China (Abecassis et al., 2025). By pooling sovereign compute resources, bridge powers can develop frontier capability while reducing global dependence on U.S. and Chinese model providers. The European Union has committed 20 billion euros through its InvestAI Facility for AI Gigafactories, with broader initiatives totalling 200 billion euros (European Commission, 2025).

This approach is not without precedent. CERN and Airbus demonstrate that like-minded nations can govern dual-use technologies through multinational structures, though AI's technical nature presents distinct challenges (Abecassis et al., 2025). Emerging pooled compute initiatives, including multilateral training projects coordinated across European and Asia-Pacific partners, demonstrate that multinational pre-training is technically and politically feasible. Early estimates suggest that access to pooled

frontier compute could allow near-frontier/ frontier-adjacent firms to "reallocate up to 70 percent of pre-training budgets to post-training, distribution, and specialized applications" (Taylor & Tan, 2025).

Canada's natural partners include EU member states (particularly France, Germany, and the Netherlands), the United Kingdom, and Asia-Pacific allies including Australia, Japan, and South Korea. Vietnam's FPT has partnered with Quebec's Mila (Abecassis et al., 2025), signalling broader interest. Canada brings distinctive assets: an AI research legacy, trusted middle power status, governance orientation, and Cohere as a potential contributor and beneficiary of partnership resources.

Open-source foundation models should also emerge as a key output of multinational partnership. In addition to avoiding lock-in, the strategic value is substantial: according to a recent Harvard study, open-source software generates approximately $2,000 in value for businesses per dollar invested (Hoffman et al., 2024). However, current open-source models carry continuity risks. Meta has reportedly pivoted toward a proprietary model codenamed "Avocado" for 2026, driven by frustration over DeepSeek using Llama architecture (Kasanmascheff, 2025). As mentioned above, even purportedly open models can impose restrictions. A multinational partnership should commit to producing more fully open models, ensuring Canada and partner nations are co-creators rather than merely consumers of the open-source models they rely upon.

IX.D Fundamental AI Research: A Sovereignty Lens

Fundamental AI research has distinct compute needs from commercial model training. From a sovereignty perspective, Canada must ensure adequate domestic research capacity for three reasons: avoiding dependency on foreign infrastructure for Canadian research, retaining talent who might otherwise migrate to better-resourced foreign institutions, and contributing meaningfully to any multinational partnership rather than serving as a passive beneficiary.

The dependency risk is very real. Canada faces a 7.6 times per capita compute disadvantage versus the United States (Dobbs & Hirsch-Allen, 2024). One estimate suggests that combining all Canadian computing resources would allow GPT-3, a model now considered very outdated, to be trained in approximately 100 days. Canada's domestic AI compute infrastructure has less than half the capacity it needs to reach parity with international competitors (Dobbs & Hirsch-Allen, 2024).

Talent retention compounds this concern. Up to 30% of technology-focused STEM graduates leave Canada, with 66% of software engineering graduates emigrating (Spicer et al., 2018). The consequences of Canada losing its AI talent are potentially more severe than with other industries, given AI's transformative potential across the economy and society (McIntyre & Golob, 2023). Countries like the United States and Singapore offer both higher salaries and superior infrastructure. Recent federal caps on international students and narrower pathways to permanent residence may further limit retention of the foreign STEM graduates Canada has been increasingly successful at keeping.

Canada's $2.4 billion Sovereign AI Compute Strategy, announced in the 2024 federal budget (Department of Finance Canada, 2024a), represents an important starting point. The Digital Research Alliance of Canada (DRAC) operates dedicated AI research clusters with $85 million in additional investment from 2025 to 2027 (DRAC, 2024). The three National AI Institutes—Mila (Montreal), Vector (Toronto), and Amii (Edmonton)—provide additional capacity. Yet these investments have not closed the gap with peer nations. Canada holds just 0.7% of global AI compute capacity, which is the lowest share among G7 nations (Dobbs & Hirsch-Allen, 2024). Demand for AI research compute in Canada also far exceeds supply. DRAC reported a 20% GPU allocation rate in 2024, meaning four out of five researcher requests went unfulfilled (DRAC, 2025). Without further investment, Canada risks falling further behind despite its early AI leadership.

If Canadian policy makers wish to change this status quo, they will need to address growing researcher demand for compute, ensure domestic capacity adequate to retain top talent, link research compute investment to multinational partnership participation, and design infrastructure for potential dual use applications.

X. STRENGTHENING STATE CAPACITY AND AI LEADERSHIP

The previous sections outlined strategic options across the AI technology stack, focusing on both inference/deployment and foundation model training. Yet strategy alone is insufficient. Building sovereign AI capacity requires institutional capacity that fits the ambition. Without effective state capacity, even well-designed policy risks falling short of implementation goals. This section examines gaps in Canada's digital and AI policy capacity and outlines options for consideration.

X.A Canada's State Capacity Challenge

Canada's approach to digital and AI governance has historically been distributed across multiple institutional homes, each with partial mandates and limited cross-government authority. The appointment of a Minister of Artificial Intelligence and Digital Innovation was an important signal of political priority. However, the broader machinery of the federal government for digital and AI remains deeply diffused and disconnected.

At minimum, six distinct institutional actors hold overlapping responsibilities for digital and AI strategy and delivery. ISED sets innovation and industrial policy. The new Minister of AI and Digital Innovation holds an economic growth mandate but inherits limited operational machinery. The Treasury Board Secretariat's (TBS) Office of the Chief Information Officer governs digital policy and standards. Shared Services Canada provides IT infrastructure. The Canadian Digital Service, which was relocated from TBS to Employment and Social Development Canada in 2023, advises departments on digital service delivery. And Public Services and Procurement Canada manages enterprise procurement and common technology solutions. This distribution reflects legitimate institutional histories and sectoral mandates, but it creates a critical structural problem: no single entity possesses both the strategic mandate and the operational authority required to drive a coherent, whole-of-government sovereign AI strategy.

This fragmentation carries measurable costs. Canada's ranking on the United Nations E-Government Development Index declined from 3rd globally in 2010 to 47th in 2024 (United Nations Department of Economic and Social Affairs [UN DESA], 2024). While multiple factors contributed to this decline, the absence of consolidated digital authority is a significant structural impediment. Fragmented governance makes prioritization difficult, slows procurement and delivery, creates gaps in technical capability, and diffuses accountability for outcomes. Section VIII.F documented how current procurement vehicles fail to distinguish between sovereignty levels and how a 30% federal IT vacancy rate constrains implementation. These are symptoms of a structural condition: without institutional consolidation, operational improvements remain incremental and uncoordinated. For a capability as strategically important as sovereign AI, these friction costs are not merely inconvenient: they risk undermining even well-resourced investments if execution authority remains dispersed across competing institutional silos.

Several peer nations have drawn a different conclusion. The United Kingdom created its Department for Science, Innovation and Technology (DSIT) in 2023, bringing digital, AI, and technology policy under a single cabinet-level department. Within DSIT, the government established a Sovereign AI Unit in 2025, backed by up to 500 million pounds in investment to scale domestic AI companies and secure strategic capabilities across the value chain (HM Government, 2025). The U.K. also appointed a dedicated Minister for AI and Online Safety with direct oversight of the AI Security Institute, semiconductor policy, and the government's Tech for Growth initiatives (PublicTechnology, 2025). Australia appointed Chief AI Officers across every federal agency, established a centralized GovAI platform for government-wide AI deployment, and anchored its National AI Plan under the Department of Industry, Science and Resources with a dedicated A$29.9 million AI Safety Institute (Australian Government, Department of Industry, Science and Resources, 2025). France installed

a National Coordinator for Artificial Intelligence reporting directly to the Prime Minister, backed by an estimated 4 billion euros in AI investment commitments (OECD, 2025).

The common thread across these models is structural clarity: consolidated authority, explicit ministerial accountability, dedicated funding attached to a single institutional home, and clear operational delivery mandates. In each case, governments recognized that distributed governance undermines execution speed and coherence, and that sovereign AI capacity requires a single institutional steward with both strategic authority and operational muscle.

X.B Strengthening AI and Digital State Capacity

If Canada intends to treat sovereign AI as a strategic national priority, the government could consider consolidating digital and AI authority into a single, empowered institutional vehicle. Two options, which are not mutually exclusive, merit consideration.

A fully resourced Ministry of Digital and AI. One path would be to create a dedicated ministry with cross-government delivery authority over digital and AI strategy. This would require two key components: a significant policy executive team capable of setting whole-of-government strategy, coordinating across departments, and owning outcomes on digital modernization and sovereign AI; and a well-staffed technical operation led by a Chief Information Officer with operational authority over digital and AI infrastructure, data governance, and delivery standards. This model follows the U.K. and Australian approaches of consolidating authority that was previously distributed across multiple departments. The transition costs, both financial and bureaucratic, would be significant, and navigating the realignment of existing institutional arrangements would require sustained political commitment. But the alternative — continuing to distribute authority across six or more actors — has produced the results reflected in Canada's digital government decline.

A dedicated Sovereign AI Unit. Whether or not broader ministerial consolidation occurs, the government could establish a dedicated Sovereign AI Unit modelled on the U.K. approach. This unit would possess explicit investment capital and a clear mandate to buy, build, and invest in major sovereign AI projects. Its mandate could encompass direct investment in Canadian sovereign cloud infrastructure (as outlined in Section VIII.B), support for Cohere and foundation model development (Section IX.B) and coordination of federal procurement to aggregate demand for sovereign public cloud (Section VIII.F), and development of sovereignty audit and certification capabilities. A Sovereign AI Unit could be implemented more quickly than full departmental consolidation and would address the most acute gap: the absence of a single entity with both funding authority and a delivery mandate for sovereign AI. It could be housed within ISED under the current framework or within a new digital ministry if Option A is pursued.

The choice of institutional model matters less than the underlying principle. Sovereign AI requires sovereign machinery. And, in government, that machinery needs a central command. Distributed governance may work adequately for routine operations and incremental improvement. But for a strategic capability that touches national security, economic competitiveness, and the future of public service delivery, fragmented authority is a structural liability that peer nations have chosen to resolve.

X.C Additional Enabling Conditions

Two additional enabling conditions could extend beyond federal institutional design to further strengthen Canada's AI and digital state capacity.

1. Digital government modernization is a core component of a sovereign AI strategy, not a parallel policy stream. Canada's decline from 3rd to 47th in the E-Government Index reflects real constraints: outdated legacy systems, incomplete cloud migration, fragmented data architectures, and a shortage of senior technical expertise in operational roles (UN DESA, 2024). While a full treatment of government digital modernization is beyond the scope of this paper, a government that has not modernized its digital foundations will struggle to deploy sovereign AI effectively, regardless of the policy framework. Anchoring digital modernization within the same institutional

home as AI strategy creates a virtuous cycle: sovereign AI infrastructure investments drive modernization, while modern digital foundations enable effective AI deployment. The U.K.'s DSIT and Australia's GovAI platform both combine AI strategy with digital government transformation for precisely this reason (Trendall, 2025; Australian Government, Department of Industry, Science and Resources, 2025).

2. Better federal-provincial-territorial coordination is a necessary condition for achieving AI sovereignty. AI sovereignty touches domains including health, education, environmental monitoring, and provincial procurement where provincial and territorial governments hold significant jurisdiction and operational authority. Tier 2 sensitive data, including health records and education data, reside primarily at the provincial level; the sovereign cloud recommendations in Section VIII.B would be substantially more effective and financially viable with provincial participation in demand aggregation. Section IX's foundation model strategy depends in part on Canadian datasets held across jurisdictions, and the cross-cutting enablers in Section VIII.F would benefit from consistent standards across all levels of government. We recommend establishing a federal-provincial-territorial working group on AI and data governance to coordinate procurement standards and data-sharing protocols, identify opportunities for shared sovereign infrastructure investment, align regulatory approaches without requiring wholesale harmonization, and support Indigenous data sovereignty principles.

XI. CONCLUSION

"We are in the midst of a rupture, not a transition." With these words at Davos on January 20, 2026, Prime Minister Mark Carney dramatically declared that the international rules-based order, on which Canada and other middle powers have long relied, had come to an end. But, he argued, the work to build what comes next is already underway. The choice lies between great power dynamics where the strong win at the expense of the weak or a new "variable geometry" of coalitions built around shared interests and shared values. Carney pledged that Canada is ready to play a leading role in helping shape this new world (Carney, 2026).

Within this changing context, artificial intelligence continues to reshape every dimension of national and economic power. AI systems now solve problems that stumped humanity for decades, drive unprecedented concentrations of private investment, and sit at the heart of great power competition between the United States and China. For Canada, a country that contributed foundationally to this technological revolution yet controls neither its commanding firms nor its critical infrastructure, the stakes could not be higher.

This paper has argued that sovereignty in the AI era means something precise: freedom from coercion. It does not mean digital isolationism, technological autarky, or retreat from the global economy. No country, not even the United States or China, can achieve complete self-sufficiency across the AI technology stack. The question for Canada is not whether to have dependencies, but how to structure them to preserve choice, reduce vulnerability to foreign leverage, and ensure that Canadian data and infrastructure remain governed by Canadian laws and values.

Our analysis examined each layer of the AI technology stack through the lens of digital sovereignty. Cloud infrastructure is the critical priority: Canada faces a compute deficit as well as extraterritorial legal exposure and even potentially service denial risks. But our country possesses capacity to address these risks through sovereign data centres built on Canadian soil and operating under Canadian law. Compute hardware presents an equally severe vulnerability that cannot be meaningfully addressed domestically; managing this dependency requires diversifying across allied suppliers, securing bilateral agreements linking Canadian

strengths to technology access, and engaging with multilateral semiconductor initiatives. Foundation model access benefits could come from procurement preferences for Canadian providers like Cohere, open-source deployment on sovereign infrastructure, and diversification across model sources. And Canada has real capacity to build strong firms at the application layer. Overall, a risk-based framework calibrates these measures, balancing key tensions between security, efficiency, innovation, and prosperity.

Canada confronts this challenge with formidable strengths. We are home to world-leading AI researchers, abundant clean energy to power the compute-intensive future, deep pools of expertise across industry, and democratic institutions worth protecting. Geoffrey Hinton's Nobel Prize and the continued contributions of Yoshua Bengio, Richard Sutton, and their students remind us that the intellectual foundations of modern AI were substantially laid on Canadian soil.

Sovereignty, not solitude. This is the premise that can guide Canada's AI strategy: maintaining the freedom to make our own choices, secure from foreign coercion, while remaining open to the partnerships and innovations that no single country can generate alone. The fracturing international order, the weaponization of technological dependencies, and the explicit framing of AI dominance as a national security imperative by our closest ally have clarified what is at stake. Middle powers must act together with shared interest and with shared values. The question is whether Canada makes the necessary choices to succeed or cedes its digital future to others.

APPENDIX A: FRAMEWORKS FOR ANALYZING CANADA'S SOVEREIGN AI STACK

This appendix presents the two analytical frameworks applied throughout the paper. Each framework serves a distinct purpose at a different stage of the analysis. The AI Technology Stack (Framework 1) provides the structure for assessing Canada's current position, identifying vulnerabilities and dependencies at each layer of the technology required to deploy AI systems. The Security and Criticality Spectrum (Framework 2 below) provides the basis for calibrating strategic options, determining which sovereignty measures are appropriate for which workloads based on sensitivity and consequence of compromise.

A.I Framework 1: The AI Technology Stack

AI systems do not exist as isolated products. They depend on a layered technology stack, from physical infrastructure at the base to end-user applications at the top. Sovereignty risks and dependencies differ at each layer. Some layers, such as compute hardware, are globally concentrated in ways Canada cannot unilaterally address. Others, such as cloud infrastructure services, present significant vulnerability but also realistic opportunities for domestic or allied alternatives.

Section VII applies this framework to assess Canada's sovereign AI capacity, examining each layer systematically to identify where Canada faces the greatest exposure and where intervention is most feasible. The assessment reveals critical vulnerabilities at the compute hardware and cloud infrastructure layers, moderate concerns at foundation models and applications, and relative strength at physical infrastructure.

The framework identifies seven key layers, ordered from foundational to applied:

Table A.1: The AI Technology Stack

Layer	Name	Description	Key Components
1	Data & Data Governance	The information that AI systems process and the rules governing its handling	Data pipelines, residency requirements, compliance frameworks (PIPEDA, sectoral regulation), vector and embedding storage
2	Physical Infrastructure & Networks	The facilities, power systems, and networks that house and connect computing resources	Power generation and distribution, data centres, secure facilities, geographic distribution, network connectivity (telco fibre, undersea cables, satellite internet)
3	Compute Hardware	Specialized processors and interconnects that execute AI workloads	GPUs, TPUs, CPUs, AI accelerators, high-speed interconnects, supply chain dependencies

Table A.1: The AI Technology Stack (cont.)

Layer	Name	Description	Key Components
4	Cloud Infrastructure Services	Virtualized infrastructure offered as managed services	IaaS/PaaS offerings, storage services, networking, managed databases, cloud control plane
5	Foundation Models	Large-scale AI models accessed for inference workloads	Proprietary model APIs (OpenAI, Anthropic), open-weight models (Llama, Mistral), Canadian alternatives (Cohere)
6	Model Operations, Serving & Orchestration	Tooling to deploy, optimize, and manage AI models in production	Inference optimization (vLLM, TensorRT), fine-tuning pipelines, MLOps platforms, orchestration frameworks, model serving infrastructure
7	Applications	End-user products and services built on AI capabilities	Enterprise applications, consumer products, domain-specific solutions, AI agents and assistants

Cross-cutting concerns: Governance and security, including encryption, access control, compliance monitoring, and audit logging, apply across all layers. Their implementation should be calibrated by the sensitivity tier described in Framework 2.

A note on scope: This stack describes infrastructure for deploying and running AI systems (inference), not for training frontier foundation models. Training infrastructure has fundamentally different requirements: massive, concentrated compute, uninterruptible power for weeks or months, and petabyte-scale data. Foundation model creation is addressed separately in Section IX of the main report.

A.II Framework 2: The Data Sensitivity Spectrum

Not all data and workloads require the same protections. Applying maximum security controls universally would be prohibitively expensive and would limit access to innovation. Applying minimal controls universally would expose critical systems to unacceptable risk. A calibrated approach requires clarity on two distinct questions: how sensitive is the data, and who holds it?

Data sensitivity determines the harm if information is compromised, ranging from grave damage to national security to minor inconvenience. Institutional context determines which governance mechanism can deliver the appropriate controls: Treasury Board directives for federal departments, collaborative frameworks for provinces, or regulation and legislation for the private sector. The framework below addresses both, organizing data into three sensitivity tiers that can be mapped across federal, provincial/municipal, and private sector contexts.

Section VIII applies this framework to calibrate strategic options, matching sovereignty requirements to both the sensitivity of the data and the institutional context that governs it.

Table A.2: Data Sensitivity Spectrum

Tier	Harm Threshold	Equivalent Federal Classification[1]	Federal Government	Provincial/Municipal Governments	Private Sector
Tier 1: Classified	Serious-to-grave harm to the **national interest** (military operations, intelligence sources, foreign relations, public safety)	Top Secret/Secret/ Confidential	Defence, intelligence, foreign affairs	*(Rare — shared classified data from the federal government)*	*(Rare — defence/ intelligence contractors with classified access)*
Tier 2: Personal/ Sensitive	Serious harm to **individuals or organizations** (identity theft, financial fraud, medical privacy violation, reputational damage, loss of competitive advantage)	Protected A/B/C	Citizen service delivery, personnel records, tax records, health admin	Health records, education, social services, vital statistics	PIPEDA-governed sensitive personal data (financial, health, biometric) Federally regulated sectors (banking, telecom, energy, transport)
Tier 3: General Operations	Limited or negligible harm (minor inconvenience, easily recoverable disruption, information already near-public)	Unclassified	Routine admin, open data, public communications, aggregate statistics	Municipal operations, open data, public notices	General business operations, consumer applications, PIPEDA-governed basic personal information (contact data, transaction records)

[1] Canada's current classification framework maps imperfectly to these tiers: notably, Protected B and Secret carry the same "serious injury" designation despite representing fundamentally different categories of harm. As Lloyd (2024) argues, reforming this framework along the lines adopted by the United Kingdom, Australia, and the United States would yield broad benefits beyond sovereignty policy. For the purposes of this paper, we map to the existing federal classifications as shown in Table A.2.

THREE TIERS OF RISK

The framework distinguishes data by the severity and nature of harm if compromised.

Tier 1 encompasses data and workloads where compromise would threaten national security, critical infrastructure operation, or public safety. This includes defence and intelligence systems, foreign affairs, and systems processing information classified at the Top Secret, Secret, or Confidential level. Even within Tier 1, dissemination caveats such as "Canadian Eyes Only" and "Five Eyes" further constrain who may access the data, creating distinct sovereignty requirements depending on whether information is restricted to Canadian personnel or shareable within allied intelligence partnerships.

Tier 2 covers sensitive information where disclosure or disruption would cause serious harm to individuals or organizations, but not at the level of national security. This includes personal information protected under federal and provincial privacy legislation, proprietary business data, and information held by federally regulated critical infrastructure sectors. The distinction from Tier 1 is not that this data is unimportant — a large-scale breach of health records or financial data can be devastating — but that the nature of the harm is different, and the appropriate controls should reflect that difference rather than defaulting to national security-grade protections.

Tier 3 covers data and workloads where harm from disclosure would be limited, such as minor financial loss, reputational inconvenience, or easily recoverable disruption. This includes unclassified government information and routine business data. For Tier 3, the primary concern is maintaining access to innovation and cost efficiency.

MAPPING ACROSS INSTITUTIONAL CONTEXTS

The same sensitivity level requires different policy levers depending on who holds the data.

Federal government data is governed directly through Treasury Board directives, the Directive on Security Management, the Privacy Act, and federal procurement policy. Ottawa can mandate sovereignty requirements for its own departments and agencies.

Provincial and municipal government data falls outside federal jurisdiction. The federal government cannot impose requirements on other levels of government. It must rely on recommendations, adoption incentives, and collaborative frameworks. Health records, education data, and social services data, all Tier 2, are held primarily at the provincial level, making intergovernmental coordination essential.

Private sector data is governed primarily through the Personal Information Protection and Electronic Documents Act (PIPEDA), market mechanisms, and voluntary standards. PIPEDA applies wherever private sector organizations hold personal information, spanning both Tier 2 (sensitive personal data such as financial, health, and biometric records, where privacy obligations are heaviest) and Tier 3 (basic personal information such as contact data and transaction records, where routine compliance applies). The distinction between the tiers is the harm threshold, not whether PIPEDA applies.

Within the private sector, **federally regulated critical infrastructure** — banking, telecommunications, energy, and transportation, designated under the Critical Cyber Systems Protection Act — occupies a distinct position. These industries are subject to sectoral regulation through OSFI, the CRTC, and the Canadian Energy Regulator, giving the federal government a regulatory lever that does not exist for the broader private sector.

BIBLIOGRAPHY

TO COME