Centre for International
Governance Innovation

Policy Brief No. 174 — June 2022

A Two-Track Approach for Trustworthy AI

Michel Girard

Key Points

- Standards and certification programs are developed to support new trustworthy artificial intelligence (AI) legislation.
- Recent developments point to the emergence of a two-track approach.
- One track would focus on the certification of AI applications embedded in tangible products using objective criteria through established conformity assessment bodies.
- Regarding AI interacting with humans and used in delivering services, there is a need to create a new track.
- This track is needed to verify and validate compliance against subjective criteria, values and ethics, seen by many as an integral part of what constitutes trustworthy AI. The practice of "assurance as a service" can be adapted to verify and validate conformance to upcoming trustworthy AI standards.

Introduction

This policy brief provides an update on key legislative and policy developments framing trustworthy AI. It sketches possible approaches for the certification, verification and validation of AI embedded in products and in services and looks at recent proposals regarding the creation of a new chartered profession to deliver assurance services to achieve trustworthy AI.

Trust is the most powerful force underlying the success of any organization — yet it can be shattered in an instant. This helps explain concerted actions by governments and industry to create a credible framework for trustworthy AI.

Technologies that use AI to react and respond in real time without human intervention are already improving business productivity. Future growth prospects are nothing short of mind boggling: a survey by management consultancy McKinsey & Company estimated that AI analytics could add US$13 trillion to annual global GDP by 2030 (Bughin et al. 2018). However, deploying AI systems comes with its share of risks. Algorithms embedded in products and equipment can trigger defects or failures, resulting in injury or death. Prejudices flowing from flawed data sets, or baked into algorithms, can harm individuals, minorities or vulnerable populations.

About the Author

Michel Girard is a senior fellow at CIGI, where he contributes expertise in the area of standards for big data and artificial intelligence (AI). His research strives to drive dialogue on what standards are, why they matter in these emerging sectors of the economy, and how to incorporate them into regulatory and procurement frameworks. He highlights issues that should be examined in the design of new technical standards governing big data and AI in order to spur innovation while also respecting privacy, security and ethical considerations.

In addition, Michel provides standardization advice to help innovative companies in their efforts to access international markets. He contributes to the CIO Strategy Council's standardization activities and advises the Chartered Professional Accountants of Canada on data governance issues.

Michel has 22 years of experience as an executive in the public and not-for-profit sectors. Prior to joining CIGI, Michel was vice president, strategy at the Standards Council of Canada, director of the Ottawa office at the Canadian Standards Association, director of international affairs at Environment Canada, corporate secretary at Agriculture Canada and acting director of education and compliance at the Canadian Environmental Assessment Agency. He holds a Ph.D. and a master's degree in history from the University of Ottawa.

Misused AI systems can lead to manipulative, exploitative and social control practices.

These risks need to be managed. According to Wael William Diab, a world expert on AI systems who helped develop a new International Organization for Standardization (ISO) standard on trustworthy AI, "Every customer — whether it's a financial services company, whether it's a retailer, whether it's a manufacturer — is going to ask: 'Who do I trust?' Many aspects including societal concerns, such as data quality, privacy, potentially unfair bias and safety must be addressed" (quoted in ISO 2020). A recent Edelman Trust Barometer report, based on annual surveys of 33,000 people, confirms that in 2021 trust in AI decreased in 25 out of 27 countries (Edelman 2021a; 2021b, 44). Consumers are concerned about the harms that AI systems can inflict because of their perceived opacity, complexity, bias and unpredictability.

As a result, there is growing demand for "trustworthy AI." Governments around the world are now taking steps to frame AI through new regulations and executive orders. As anticipated, compliance to mandatory requirements will rely on digital governance standards, certification programs and accreditation schemes (Girard 2019). Governments incorporate standards in a wide range of health and safety regulations. Those regulating AI have signalled their intention to take the same approach.

However, standardizing trustworthy AI will be a complex task. AI chips, algorithms and machine learning can be embedded in virtually any product, system or service. Although standards, testing and certification work well when objective criteria are assessed, standardizing highly subjective criteria such as values, ethics or trust is greenfield territory. Policy makers appear to be moving toward the creation of a two-track approach. One track would focus on the certification of AI applications embedded in tangible products using established conformity assessment bodies. Regarding AI interacting with humans and used in delivering services, industry and governments in Europe are looking to create a new track for the verification and validation of AI systems. This track could be based on international management system standards such as ISO 17021 or the recently released ISO 17029, an international accreditation standard allowing for the use of chartered professionals to perform service engagements

supporting digital governance standards, including AI (Standards Council of Canada 2022).

Laws, Regulations and Executive Orders

Governments around the world are looking at making AI trustworthy through standards. As expected, the European Union, the United Kingdom and China have pledged to incorporate international digital governance standards and certification programs as a compliance mechanism in upcoming regulations. In the United States, executive orders issued by the White House are calling for the creation of voluntary standards for trustworthy AI that should be adopted by federal departments and agencies and by industry.

The EU Artificial Intelligence Act tabled in 2021 states that developers and users of high-risk AI will need to abide by international standards and certification programs. AI regulations will frame high-risk AI applications in the delivery of products, devices, systems, networks and services. At the organization level, the European Union will require the adoption of enterprise-wide quality management and risk management system standards for any organization developing or using high-risk AI. Organizational compliance to upcoming AI risk management standards will be audited by independent third parties. In addition, AI embedded in regulated consumer products and in machines has been defined as high risk and will need to meet new safety standards. Standards will also be developed to cover predefined high-risk AI systems interacting with humans and delivering services to consumers and citizens.[1]

The European Union has created a Supervisory Agency of AI that will manage the implementation of the AI strategy and the upcoming regulatory framework for AI.[2] The agency is now working on crafting AI regulations and will determine how best to incorporate standards and certification programs as a compliance mechanism. AI standards are expected to play a large role in shaping what it means to adhere to the proposed conformity assessment regime.[3]

The Government of the United Kingdom (2021) recently announced it will implement a high-risk AI regime that should be equivalent to the EU AI Act. The United Kingdom will regulate high-risk AI through standards and certification programs. In December 2021, the government signalled its intention to become a world leader in the development of international digital governance standards and certification programs through the creation of an AI Standards Hub (Government of the United Kingdom 2022).

The Government of China has also signalled its intention to use standards as a compliance mechanism for managing "high-stakes" AI systems. Building on the 2017 AI Development Plan, government agencies rolled out a wide array of AI governance policies and frameworks in 2021. While China operates in a dramatically different political environment, the proposed approaches touch on issues of shared concern, such as algorithmic transparency for users and the testing and certification of high-stakes AI systems. Next steps include the development of a trustworthy AI framework based on a comprehensive review of AI ethical constraints, normative legislation and best practices to serve as a set of methodologies for implementing AI governance requirements (Ernst 2020). It is expected that standards and certification will play a central role in framing trustworthy AI in China as the country continues to enhance its participation in international standards development bodies and technical committees focusing on digital governance. Its recently released "China Standards 2035" strategy reaffirms this strategic thrust (Koty 2020).

The US government has taken a different approach by avoiding imposing regulatory constraints on the AI industry, focusing instead on the development of voluntary standards and certification to achieve trustworthy AI. In response to an executive order from the White House, the US National Institute for Standards and Technology (NIST) set in motion the "US Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools." Through the plan, voluntary

1 *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, [2021] COM/2021/206 final, online: <https://perma.cc/H42G-AB3Q>.

2 See https://digital-strategy.ec.europa.eu/en/policies/plan-ai.

3 See www.tuev-verband.de/digitalisierung/kuenstliche-intelligenz.

AI standards and certification programs would be developed to support AI risk management frameworks for organizations. NIST would also facilitate the creation of a national standards framework for trustworthy AI systems. Industry will be called upon to participate in standards development activities with standards bodies nationally, regionally and internationally. Once developed, standards and certification programs would be used by federal departments and agencies in procurement activities to encourage their widespread adoption (NIST 2019). Looking forward, this approach may change somewhat as legislators have recently tabled draft legislation to regulate digital platforms including algorithmic processes to ensure they are fair, transparent and safe. This would be undertaken by a "Code Council," which "shall develop proposed voluntary or enforceable behavioral codes, technical standards, or other policies for digital platforms" to govern their function.[4]

AI in Tangible Products and Systems

As indicated above, governments have decided to use standards as a compliance mechanism to ensure that AI-enabled products remain safe under normal operations and extreme conditions. It calls for the development of common normative standards for all high-risk AI systems. AI systems used as components of products already standardized to meet safety requirements have been labelled as high-risk AI "in order to ensure that only safe and otherwise compliant products find their way into the market" (European Commission 2021, 24). The draft regulation notes that "the safety risks that may be generated by a product as a whole due to its digital components including AI systems should be prevented and mitigated" (ibid.).

The regulation's coverage is expected to be extensive. It will include AI components deployed in any electrical, plumbing, gas and heating/cooling systems currently covered by safety regulations. Toys, recreational equipment, systems intended for use in explosive atmosphere and cableway installations are also covered, along with systems embedded in infrastructure such as elevating devices, boilers and pressure vessels, waterworks, machinery, radio equipment and telecommunications. The regulation also targets critical infrastructure not currently standardized through safety codes (ibid., 25).

It is expected that new AI safety standards focusing on the performance of AI chips and algorithms embedded in tangible products and systems will be added to product safety standards and safety codes currently in force. Accredited conformity assessment bodies managing the certification of products and systems are well positioned to deliver these additional services through long-standing relationships with product manufacturers. In order to reduce barriers to trade, it is expected that testing and certification of AI components will be performed under relevant ISO accreditation standards, notably ISO 17065 for the certification of products, and ISO 17025 for testing and calibration laboratories. This allows national testing results to be recognized between jurisdictions, thereby avoiding multiple testing requirements when exporting products abroad.

Although specific safety standards for AI embedded in products and systems have not been developed yet, preliminary work has begun. On that front, NIST recently started to articulate objective and measurable criteria that could be used for the testing and certification of safe AI in products and machines (NIST 2020). Although this is a work in progress, it could provide guidance on how to achieve outcomes such as:

- accuracy (AI systems should make the right decisions);
- reliability (AI systems should be designed to operate continuously and consistently);
- resiliency (when new data causes an AI system to operate outside of its nominal boundaries, it should be able to adapt to new conditions or to alert humans in order to avoid catastrophic failure);
- safety (AI systems should not create health or safety hazards to humans or the environment); and
- explainability (the process used by AI systems to make decisions should be

4 US, Bill S, *A Bill to establish a new Federal body to provide reasonable oversight and regulation of digital platforms*, 117th Cong, 2022.

documented, understood and replicated by humans — this requires transparency).

One could envisage performance standards providing guidance to achieve these outcomes for various categories of products and systems, as well as specific testing procedures to certify that AI applications embedded in products and machines are safe.

AI in Services

Clearly, all AI systems interacting with humans and providing services to consumers or citizens are not created equal. Some, such as algorithms that propose music playlists and movie suggestions through streaming services, or promote products through online advertising, can probably be managed without too much concern about harm creation.

However, AI systems that can breach the fundamental rights of persons are considered high-risk AI. The EU regulation lists a series of high-risk services or interactions. They include real-time and post remote biometric identification systems; systems used in education and vocational training; AI systems testing persons as part of their education; AI systems used in employment, worker recruitment and management; systems used to evaluate credit scores and managing eligibility to essential services such as finance, insurance, housing, electricity and telecommunications; as well as systems used for law enforcement and the administration of justice (European Commission 2021, 26–27).

In addition to objective criteria applied to products and machines, standards framing trustworthy AI used in services will likely assess compliance to subjective criteria, including values and ethics. These values are seen by many as an integral part of what constitutes trustworthy AI. Many statements and declarations on ethical and trustworthy AI make reference to these values.[5] Subjective values include fairness, equity and privacy. AI systems would have to respect the rights of individuals enshrined in laws and regulations and be devoid of prejudice or bias against individuals or groups. They may also have to support broader objectives such as inclusive growth, sustainable development and well-being and contribute to beneficial outcomes for people and the planet.

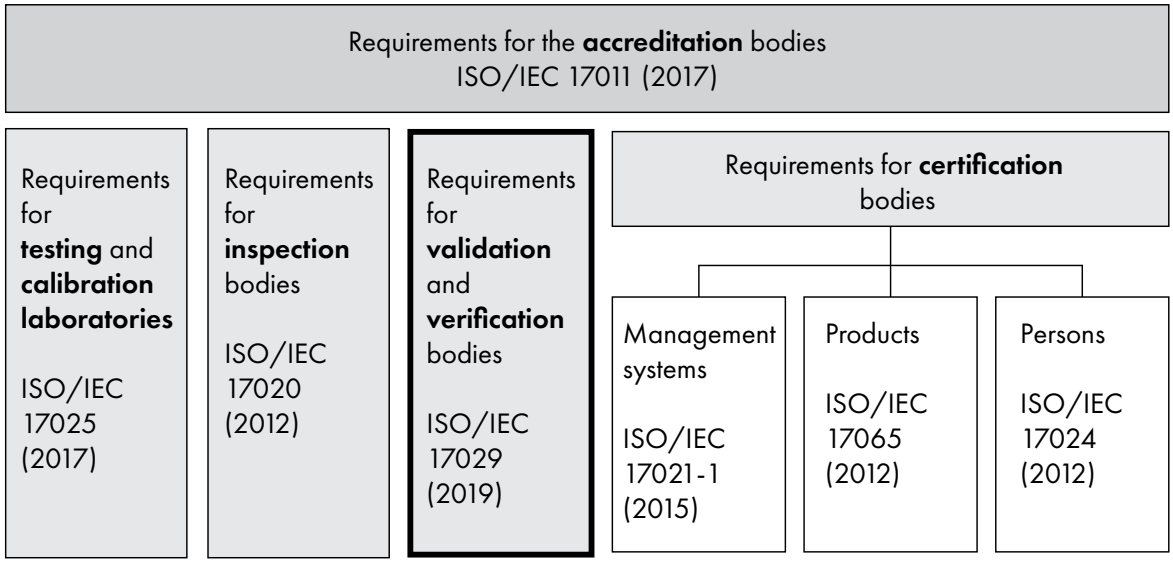
Governments and industry generally recognize that compliance to subjective values cannot be demonstrated through tests in a laboratory. It requires human interaction and judgment. This explains why the European Union and the United Kingdom are now exploring the creation of what could be described as a separate track to verify and validate AI systems interacting directly with humans. This new track could be based on the recently released ISO 17029 accreditation standard entitled “Requirements for Validation and Verification Bodies” (ISO/IEC 2019). This new standard allows for claims by organizations adhering to new digital governance standards to be verified by independent third parties.

As Figure 1 shows, ISO 17029 is complementary to other ISO accreditation standards. It avoids duplication by focusing on activities not covered by other 17000 series standards. Activities including industrial automation systems, software and systems engineering, AI and information technologies can now be verified and validated through programs developed under ISO 17029. Validation/verification programs are defined as a set of rules, procedures and management for carrying out validation/verification activities in a specific sector or field, specifying the scope of validation/verification, competence criteria, process steps, evidence-gathering activities and reporting (Committee on Conformity Assessment 2019).

The release of ISO 17029 in 2019 represented a watershed moment in the international conformity assessment ecosystem because it formally opened the door to professional classes, such as

5 Among the many declarations and statements on this issue, one notes the Asilomar Principles proposed by the Future of Life Institute (https://futureoflife.org/ai-principles/); the Open Data Charter (https://opendatacharter.net/principles/); the 2017 Montréal Declaration for a Responsible Development of Artificial Intelligence (www.montrealdeclaration-responsibleai.com/the-declaration); and the Top 10 Principles for Ethical Artificial Intelligence (www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf).

Figure 1: ISO/IEC Accreditation Standards



Source: Committee on Conformity Assessment (2019).

chartered accountants, engineers or auditors, to perform conformity assessment work against value-laden digital governance standards.

That being said, two challenges will have to be overcome. First, ISO 17029 is a newly released standard and has not been widely adopted yet. International accreditation organizations are still exploring options for the appropriate recognition model for AI systems in order to achieve mutual recognition between jurisdictions. ISO/IEC 17021-1, another well-established accreditation standard focusing on management systems, could provide expediency and speed in achieving the same outcome.

Second, it is not yet clear whether any chartered professional class is able and willing to perform these tasks. Governments and industry will need to engage with chartered professional organizations to determine whether existing professions can acquire new competencies to provide this service, or whether an entirely new profession needs to be created. On that front, the UK government has taken a leading role by launching a process to build what has been termed "an effective, mature assurance ecosystem for AI" (Centre for Data Ethics and Innovation [CDEI] 2021). In December 2021, it unveiled its "Roadmap to an Effective AI Assurance Ecosystem." Developed by the CDEI, the road map proposes steps to build such an ecosystem in order to manage risks associated with the deployment of AI in products, operations, processes, systems and networks and generate trustworthiness to both operators and customers (ibid.).

The CDEI proposes to adapt the practice of "assurance as a service" as a platform to perform validation and verification services against upcoming trustworthy AI standards. Assurance as a service enables people to assess whether systems are trustworthy. It originates from the accounting profession, and is used by chartered professionals to cover many domains, such as quality management and cybersecurity, using international standards issued by these professions. The road map proposes to replicate what has been achieved in these mature ecosystems with digital governance standards.

The CDEI proposes five principles of assurance products and services, which are drawn from the accounting profession:

- a three-party relationship (composed of a responsible party, a practitioner and an assurance user);
- agreed and appropriate subject matter (the information can be subjected to procedures for gathering sufficient and appropriate evidence);
- suitable criteria (required for the consistent measurement and evaluation

of the subject matter within the context of professional judgment);

- sufficient and appropriate evidence (sufficiency relates to the quantity of evidence whereas appropriateness relates to the quality of the evidence, its relevance and reliability); and
- conclusions (the assurance obtained about the subject matter information).

Although accountants could play an important role in delivering assurance engagements, the CDEI notes that similar roles, responsibilities and institutions for standard setting, assessment and verification are present across the range of assurance ecosystems — from cybersecurity to product safety — providing transferable assurance approaches. As such, it is looking at the creation of new AI assurance professionals across sectors. The components required to accredit professionals include courses, degrees and vocational programs that provide qualifications for developers or assurance practitioners to formal accreditation by chartered bodies demonstrating clear professional standards.

In the United Kingdom, bodies such as UKAS (the United Kingdom's national accreditation body); BCS (The Chartered Institute for IT, formerly the British Computer Society) and the ICAEW (the Institute for Chartered Accountants in England and Wales) have been approached and will be consulted on the creation of a suitable framework for individual accreditation. The goal is to assess the most promising routes to professionalizing AI assurance across domains. Given the fact that AI is a multipurpose technology that is being deployed across all sectors of the economy, developing routes to professionalization will require coordination between a number of different actors.

Considerations

Although the United States and China are making progress in standardizing trustworthy AI, the European approach appears to be the most likely to succeed in becoming the international reference that all jurisdictions will want to emulate. Encouraged by its past success in making the General Data Protection Regulation the de facto global privacy benchmark, the European Union is now aiming to repeat the feat with AI.

The implications for its trading partners are significant. When it comes to products, machines and infrastructure, the European Union and the United Kingdom appear aligned in their quest to modernize a huge number of standards and safety codes by adding trustworthy AI requirements. Moving forward, Canada and nations trading with the European Union will have to carefully consider how best to incorporate trustworthy AI requirements into their own standards, safety codes and regulations.

Regarding AI systems interacting with humans and delivering services, Canada and other developed nations also need to engage with chartered professionals on how best to deliver assurance as a service to digital governance standards, including AI. Ideally, a consensus can be reached on common competency profiles, training and personnel certification requirements that can then be adopted by various chartered professional associations.

In order to gain an equivalency status with the European Union, developed economies such as Canada would be well served by implementing credible policies, regulations and standards that meet or exceed EU requirements. This is bound to benefit consumers/citizens and facilitate the trade of trustworthy AI systems around the world.

Works Cited

Bughin, Jacques, Jeongmin Seong, James Manyika, Michael Chui and Raoul Joshi. 2018. "Notes from the AI frontier: modeling the impact of AI on the world economy." Discussion Paper. McKinsey Global Institute. September. www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world-economy.

CDEI. 2021. "The roadmap to an effective AI assurance ecosystem — extended version." Government of the United Kingdom. December 8. www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem/the-roadmap-to-an-effective-ai-assurance-ecosystem-extended-version.

Committee on Conformity Assessment. 2019. "ISO/IEC 17029: Conformity assessment — General principles and requirements for validation and verification bodies." 34th Plenary Meeting, Nairobi, Kenya, May 1–2. https://european-accreditation.org/wp-content/uploads/2019/06/1.-S.Vehring-ISO-IEC-17029.pdf.

Edelman. 2021a. "Edelman Trust Barometer 2021: Global Report." www.edelman.com/sites/g/files/aatuss191/files/2021-03/2021%20Edelman%20Trust%20Barometer.pdf.

———. 2021b. "Edelman Trust Barometer 2021: Global Report — Trust in Technology." www.edelman.com/sites/g/files/aatuss191/files/2021-03/2021%20Edelman%20Trust%20Barometer%20Tech%20Sector%20Report_0.pdf.

Ernst, Dieter. 2020. *Competing in Artificial Intelligence Chips: China's Challenge amid Technology War.* CIGI Special Report. Waterloo, ON: CIGI. www.cigionline.org/publications/competing-artificial-intelligence-chips-chinas-challenge-amid-technology-war/.

Girard, Michel. 2019. *Big Data Analytics Need Standards to Thrive: What Standards Are and Why They Matter.* CIGI Paper No. 209. Waterloo, ON: CIGI. www.cigionline.org/publications/big-data-analytics-need-standards-thrive-what-standards-are-and-why-they-matter/.

Government of the United Kingdom. 2021. "National AI Strategy." September. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1020402/National_AI_Strategy_-_PDF_version.pdf.

———. 2022. "New UK initiative to shape global standards for Artificial Intelligence." Press release, January 12. www.gov.uk/government/news/new-uk-initiative-to-shape-global-standards-for-artificial-intelligence.

ISO. 2020. "Towards a Trustworthy AI." July. www.iso.org/news/ref2530.html.

ISO/IEC. 2019. "Conformity assessment — General principles and requirements for validation and verification bodies." 1st ed. October.

Koty, Alexander Chipman. 2020. "What is the China Standards 2035 Plan and How Will it Impact Emerging Industries?" China Briefing, July 2. www.china-briefing.com/news/what-is-china-standards-2035-plan-how-will-it-impact-emerging-technologies-what-is-link-made-in-china-2025-goals/.

NIST. 2019. "US Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools." NIST, August. www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf.

———. 2020. "Trustworthy AI: A Q&A With NIST's Chuck Romine." *Taking Measure* (blog), January 21. www.nist.gov/blogs/taking-measure/trustworthy-ai-qa-nists-chuck-romine.

Standards Council of Canada. 2022. "ISO/IEC DIS 17029, Conformity Assessment: General Principles and Requirements for Validation and Verification Bodies." Online Browsing Platform. March 17. https://scc.isolutions.iso.org/obp/ui#iso:std:iso-iec:17029:dis:ed-1:v1:en.

About CIGI

The Centre for International Governance Innovation (CIGI) is an independent, non-partisan think tank whose peer-reviewed research and trusted analysis influence policy makers to innovate. Our global network of multidisciplinary researchers and strategic partnerships provide policy solutions for the digital era with one goal: to improve people's lives everywhere. Headquartered in Waterloo, Canada, CIGI has received support from the Government of Canada, the Government of Ontario and founder Jim Balsillie.

À propos du CIGI

Le Centre pour l'innovation dans la gouvernance internationale (CIGI) est un groupe de réflexion indépendant et non partisan dont les recherches évaluées par des pairs et les analyses fiables incitent les décideurs à innover. Grâce à son réseau mondial de chercheurs pluridisciplinaires et de partenariats stratégiques, le CIGI offre des solutions politiques adaptées à l'ère numérique dans le seul but d'améliorer la vie des gens du monde entier. Le CIGI, dont le siège se trouve à Waterloo, au Canada, bénéficie du soutien du gouvernement du Canada, du gouvernement de l'Ontario et de son fondateur, Jim Balsillie.

Credits

Managing Director of Digital Economy Robert Fay
Publications Editor Jennifer Goyder
Graphic Designer Brooklynn Schwartz

For publications enquiries, please contact publications@cigionline.org.

Printed in Canada on Forest Stewardship Council® certified paper containing 100% post-consumer fibre.

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org