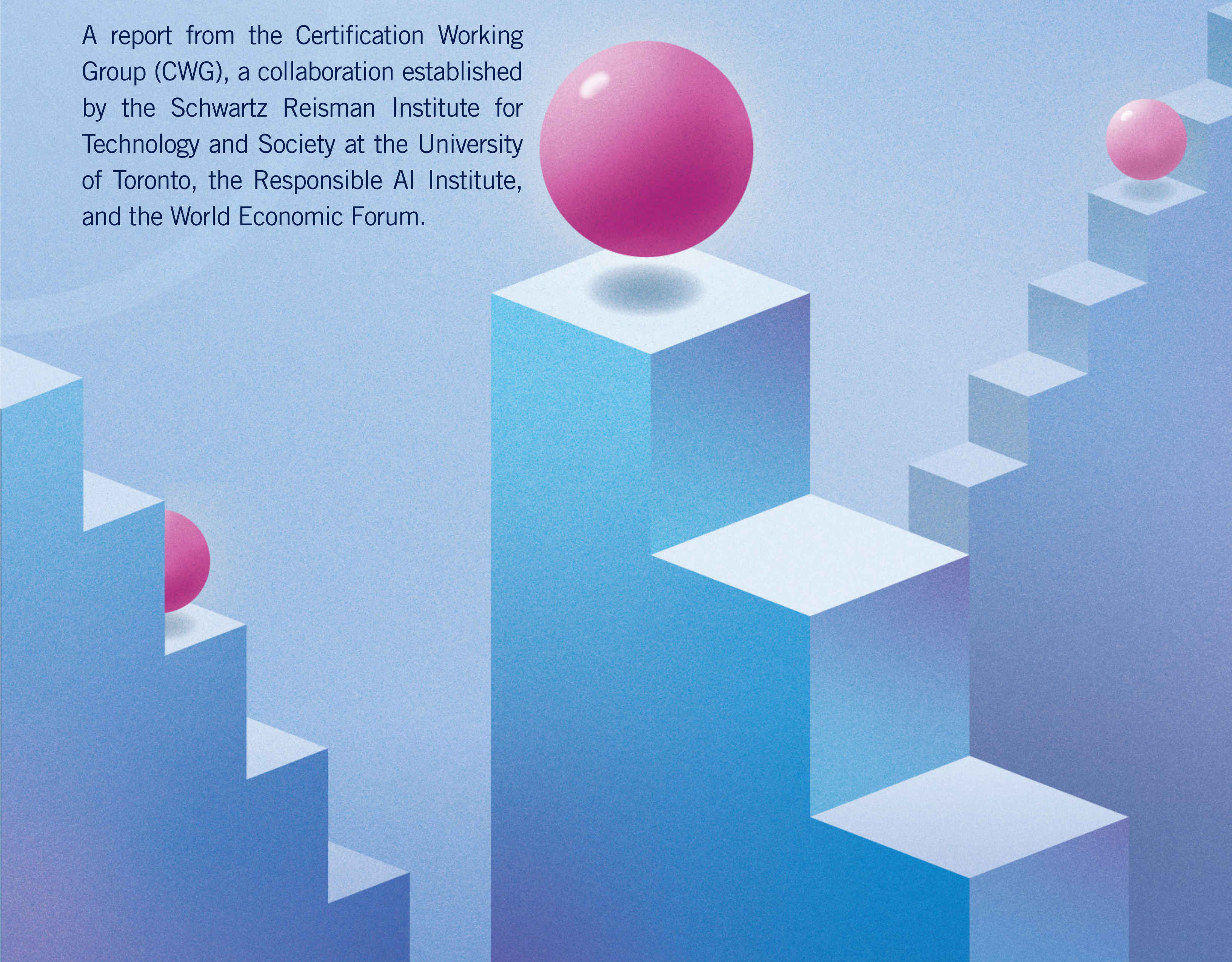
UNIVERSITY OF TORONTO

SCHWARTZ REISMAN INSTITUTE FOR TECHNOLOGY AND SOCIETY

ARTIFICIAL INTELLIGENCE CERTIFICATION

Unlocking the power of AI through innovation and trust

A report from the Certification Working Group (CWG), a collaboration established by the Schwartz Reisman Institute for Technology and Society at the University of Toronto, the Responsible AI Institute, and the World Economic Forum.



CONTRIBUTORS

Gillian Hadfield
Maggie Arai
Schwartz Reisman Institute
for Technology and Society,
University of Toronto

Ashley Casovan
Var Shankar
Responsible AI Institute

Jayant Narayan
UN Consultant - Advisor

Ron Bodkin
ChainML

Cathy Cobey
Yvonne Zhu
Ernst & Young

Phil Dawson
Armilla.ai

Jesslyn Dymond
Lindsay Green-Noble
TELUS

Gemma Galdón Clavell
Eticas Research and Consulting

Heidy Khlaaf
Trail of Bits

Brent Mittelstadt
Oxford Internet Institute

Alexander Scott
Dominique Payette
Borealis AI

Alexey Rubtsov
Global Risk Institute

Julia Stoyanovich
New York University

Craig Shank
CES.World PLLC
Chair, CWG

Note: the views expressed in this paper represent outcomes from discussions and debates among the individuals engaged with the Certification Working Group. The views expressed in this paper do not necessarily represent the positions and views of any of their employers.

TABLE OF CONTENTS

This report is published by the Schwartz Reisman Institute for Technology and Society at the University of Toronto.
Date of publication: April 2, 2024. For media inquiries, please contact sri.communications@utoronto.ca.

Executive summary

Recent developments have created new visibility into the power, potential, and risks presented by ongoing advancement in AI. Incremental progress in the development of generative AI hit a crucial tipping point in late 2022 and early 2023 with improvements in large language models (LLMs). Some LLMs, such as ChatGPT, became household names, with easy-to-use interfaces and seemingly-magical capabilities. But it soon became clear that LLMs can "hallucinate" facts, generate biased outputs, inspire unwarranted confidence among their users, and pose context-specific challenges best understood by practitioners and experts. The newfound adoption and prominence of LLMs has elevated the AI governance conversation, bringing the issue to the top of every CEO and policymaker's agenda. These events and visibility have driven important urgency into debates about whether and how to govern AI in all its forms. G7 leaders meeting in Japan in May 2023 agreed to "advance international discussions on inclusive artificial intelligence (AI) governance and interoperability to achieve our common vision and goal of trustworthy AI."[1]

There is much to gain from these new technologies. Promoting innovation and gaining the corresponding economic and societal value—while protecting us from potential harms—will be among the crucial policy balancing acts of our time.

Certification of AI systems is one cornerstone necessary to achieve policy objectives and build the necessary trust in AI. But the work to establish certification mechanisms for AI has received limited attention in policy and academic circles.

The insights and recommendations in this paper come from two years of work by the Certification Working Group (CWG), a multinational, interdisciplinary group of experts that brings together key voices with academic, government, NGO, and corporate backgrounds in emerging technologies, law and policy, governance, evaluation, engineering, audits, standards, and certification. This paper draws on these voices and highlights the need for an effective certification ecosystem for AI and what is required for it to succeed. Key features are already in place, such as trusted certification bodies, digital technology providers, and a vibrant ecosystem of startups. However, there remain gaps that must be addressed to ensure a robust certification ecosystem. These include uncertainty about the readiness of upcoming standards for AI, limited demand signal in the marketplace, and minimal investment in ecosystem development. Closing these gaps will be fundamental to building certification as a platform for innovation and for technologies that earn trust.

The principles and recommendations outlined in this report are intended to be global—a scope that is at the core of so-called "soft law" governance models that rely on mechanisms such as standards, certification, third party audits, and other independent assurance techniques. While this report makes references to specific national bodies like the United States' Federal Trade Commission, these are intended to serve as examples whose implications may be useful to an international readership.

In this report, we put forward the following recommendations to government, academic, private, and civil society stakeholders:

Government

- Lead the way in establishing **fundamental objectives** for AI certification standards and certifications. This includes investing in internal capabilities and workforces to support and foster experts who understand conformity assessment and certifications, particularly in the context of AI.

- Support **market development and demand signal** for AI certification. Ideally this should be government-led, but private organizations with large procurement budgets may also assist with this push.

All Stakeholders

- Invest **time and resources** to get the **foundations** in place, such as developing a viable first set of internationally recognized documents to support certification, clarifying what frameworks can be used for conformity assessment, and enabling joint certifications.
- Collaborate to develop an effective **AI reference architecture** for policy and accountability, enabling clarification of roles and responsibilities (including shared responsibilities), exchange of documentation between suppliers at different points in the AI ecosystem necessary to deliver a specific implementation, and maintenance of the chain of accountability between participants.
- Move quickly to **advance the state of certification** from these foundations: e.g. build transparency and data availability into next generation standards, advance environmental, social, and governance (ESG) uses for AI certification tools, and develop a focused research agenda to support continued advancement in AI verification and validation tools to improve certification.

> Promoting innovation and gaining the corresponding economic and societal value—while protecting us from potential harms—will be among the crucial policy balancing acts of our time.

Background

The Certification Working Group

The Certification Working Group (CWG) is a multinational, interdisciplinary group of experts with academic, government, NGO, and corporate backgrounds in emerging technologies, law and policy, governance, evaluation, engineering, audits, standards, and certification. Launched by the Schwartz Reisman Institute for Technology and Society at the University of Toronto, the Responsible AI Institute, and the World Economic Forum's Centre for the Fourth Industrial Revolution, CWG aims to foster the development of certification (and related certification marks) as recognized frameworks that validate AI tools and technologies as responsible, trustworthy, ethical, and fair.

Purpose of this report

This paper aims to capture key themes from over a year of interviews and small-group conversations centered on certification as a trust mechanism for AI. As a multinational, interdisciplinary group, CWG aims for this paper to provide recommendations and input that are useful to those seeking to advance governance, risk management, and trust for AI. The paper is intended to be useful to those in government, industry, civil society, and academia across the globe who see value in, and wish to advance, assurance ecosystems to ensure the responsible use of AI.

Development process

Beginning in 2021, CWG conducted research and held a series of small-group conversations with key contributors to AI, AI governance, and the certification, assurance, and regulatory technologies ecosystem for AI from around the world. These have included representatives from academia with expertise on fairness, bias, and governance; leading early-stage and venture-backed AI governance companies; leading AI developers; and large tech companies. These conversations continue to provide CWG with ongoing inputs to strategies aimed at developing and promoting effective certification approaches for AI.

Introduction

Rapid developments in AI and other digital technologies bring us into a new technology revolution that promises to transform every field of human endeavour.

The benefits could be enormous. But it is also possible to look at the same technologies and wonder where they may take us. People are rightly concerned about what this means for their jobs, their information, and their privacy. As we rely more and more on automation to make decisions, we see important and difficult issues, often touching on safety, privacy, equity, our democratic institutions, the value of our work as individuals—and over time, even what it means to be human.

Society stands to gain—both economically and in a myriad of other ways—if innovation in AI advances in ways that earn trust. Clear trust mechanisms and visible boundaries create a foundation for effective use of new technologies and a platform for innovation across economies. Building trust in AI will demand frameworks of laws and agreements crafted by governments and shaped by open discussion among all who have a stake in the outcome: citizens, business leaders, and academics, to name a few. Developing trust in technologies isn't for regulators and legislators alone. The pace and breadth of AI development will require additional, complementary tools. Markets and regulatory frameworks are not mature enough to handle the scale of deployment necessary to mandate, evaluate, or verify each element of trusted AI, nor to deliver on the necessary enforcement. Moreover, legality alone does not necessarily lead to widespread trust. How do we feel about payday lending, robocalling, certain used car lots—or even our favourite social network? Each of these legal behaviours comes with its own source of public distrust. Advanced technologies, too, have sources of distrust beyond the law, including rapid change, opaque methods, and uncertain objectives. These technologies will need to earn trust in ways that go beyond traditional law. This turns our attention to other fundamental trust mechanisms.

It's easy for us to start our morning trusting a variety of things: the light switch in our bedroom, the coffee made by a stranger at a coffee shop, the car we drive or the public transit we take. Considering these through their entire chain shows how we may have trusted thousands of people on our way out the door today. Why did we trust them? Trust in the products, processes, and technologies we encounter stems from a combination of factors including regulation, industry certifications, accreditation, and standards. While there aren't laws governing every facet of the chains that deliver products and services, we at least know that, say, the person who handles food at a restaurant has undergone food safety training and accreditation, ensuring their compliance with rigorous standards and protocols. Each industrial revolution has brought new technologies into the world that need to earn trust…

A big part of the foundational trust in technologies comes from the thousands of connected, repeatable interactions that may be audited, certified, or otherwise confirmed. Though regulation has a key role to play in these processes, a truly effective ecosystem is one which utilizes the range of trust and transparency tools available.

Our focus here is development of the necessary elements for an ecosystem that can reliably deliver effective certification to support AI that is responsible, trustworthy, ethical, and fair. By certification, we mean a process through which an independent body attests that an organization or its personnel, systems, or products meet specified objective standards or requirements, typically through the issuance of a "mark" or "label." An effective AI certification system can support societal expectations and advance innovation by both building trust in technologies and monitoring the "fences" intended to contain them.

The AI certification ecosystem: almost ready but not running yet

What is an AI certification/assurance ecosystem?

Certification and assurance ecosystems are typically market-based, often with government oversight. They don't happen by serendipity. Key participants required for such an ecosystem—like developers and certifiers—will only invest if they see that necessary market power, technical credibility (including evaluation techniques), legitimacy, and above all demand, are likely to come together. Market signal to those participants is critical to accelerating their steps in building the AI assurance ecosystem. Developing regulation plays a role as well, and may explicitly require (or at least favour) systems that are certified against trusted standards.[2] The EU AI Act in particular anticipates—and depends on—a well-developed AI certification ecosystem. But much remains to be done.

None of this investment happens without a reasonable market signal that the certification has generated enough demand that some form of return is likely for the participants.

A typical assurance ecosystem relies on third-party assurance providers (for example, companies offering audit or certification services). These third parties provide vital assessment, testing, and verification services to build genuine and justifiable trust in AI, for consumers purchasing new products and organizations procuring new AI systems alike. The validity of most certification systems also rests on some level of government oversight. For example, to offer legitimate certification services, a certification body will typically need to be accredited by an accreditation body (e.g., the American National Standards Institute's National Accreditation Board, the Standards Council of Canada, or the United Kingdom Accreditation Service). Standards also form an important part of this ecosystem, by providing a guiding framework against which certification services can be developed.

Investment by private parties is critical to the development of a certification ecosystem. Industry participants and NGOs invest time and effort in standards development. Third parties seeking to provide certifications invest substantial amounts in the effort to build on the resulting standards by creating the necessary certification schemes, tools, and internal training and oversight. Organizations hoping to receive a certification invest in their internal systems and organization of data to support the necessary audits. Yet none of this investment happens without a reasonable market signal that the certification has generated enough demand that some form of return is likely for the participants.

Based on our investigations at CWG, we believe many of the key attributes needed for an effective AI certification ecosystem are already in place. Trusted certification bodies for complex IT are well-developed and include a robust community of both small local businesses and large multinational companies. Digital technology providers are highly familiar with what it takes to comply with customer-critical certification requirements in areas like security and privacy, and AI requirements will build on those. Many areas (like safety-critical software standards under the International Electrotechnical Commission's international standard 61508[3]) have a long history of delivering effective outcomes through structured risk assessments, even in scenarios presenting problems that may be "unknown unknowns." AI standards that could serve certification are starting to reach publication milestones. Moreover, a vibrant ecosystem of regulatory technology startups has developed that can track steps in governance and controls compliance for a wide range of AI use cases and an equally wide range of requirements (whether from standards, regulations, or vertical industry norms).

The expansive impact and public visibility associated with generative AI tools bring AI's trustworthiness to the forefront of the public conversation and policy agenda. Governance and certification fundamentals still work with generative AI, but require wide, well-understood adoption, rigor in application, and a strong focus on risk assessment in wide areas of impact and potential harm.

The Certification Landscape

Some examples of types of organizations and the roles these parties play in the certification landscape:

Regulatory/Standards Landscape	International Guidelines, global laws/guidelines, national regulations, state legislation, sector regulations, best practice guidelines, proposed legislation, standards, etc	FDA Health Canada Santé Canada
Emerging Best Practices	Industry and international organization recommendations	CLEAR Derm Consensus Guidelines DATA NUTRITION PROJECT
Conformity Assessment Schemes	A description of the specific requirements, objects, and methodology	AI Responsible Artificial Intelligence Institute Advancing Trusted AI ISO ISO/IEC
Accreditation	Demonstration of its competence, impartiality and consistent operation in performing specific conformity assessment activities	ANSI American National Standards Institute SCC CCN
Testing and Evaluation	Determination of one or more characteristics of an object of conformity assessment according to a procedure	ARMILLA FAIRLY credo ai
Audit	Obtaining information about an object of conformity assessment and evaluating it to determine the extent to which specified requirements are fulfilled	bsi. intertek UL
Certification	Third–party attestation related to an object of conformity assessment, indicating that it met specific requirements	MDSAP MEDICAL DEVICE SINGLE AUDIT PROGRAM

Where are the gaps in the current assurance ecosystem?

At the outset, AI certification faces a challenge that boils down to two key questions. First, what objectives are we trying to achieve through certification? And second, who decides what those objectives should be? We can see this in early drafts of management systems and governance standards for AI—which either do not specify (by design) the core meaning of "responsible" AI development,[4] or specify it at a very detailed level that is not necessarily consistent with governmental or societal objectives, though the aspirations may be laudable.[5]

The level at which these standards currently operate is very much centered on the systems for making and implementing decisions. These are important steps to standardize, yet they lack clear direction on what to solve for: is it safety? Trustworthiness? Efficiency? Profit? There will be more development in these standards that will aim to add clarity through related documents and practices. However, the standards system is not the appropriate venue for necessary conversations about what the real objectives should be for responsible behaviour for AI developers and implementers. In the end, the real cornerstone needed is a meaningful expression of societal values, expectations, and rules, established through legitimate governmental (or intergovernmental) processes. The regulatory participants that hope to rely on certification and other assurance mechanisms need to play a "first chair" role in developing clear direction about what values responsible AI must live up to, and what certification is meant to achieve.

While government has the necessary credibility, legitimacy, and authority to provide direction for responsible AI, experts who understand both AI and the role certification (and supporting standards) can play are too few and too far between, and existing experts will be stretched too thin to cover expanding needs. Without an adequate workforce trained in these areas, or a budget to hire and train them, it will be difficult for government to ensure that the right values drive key AI certification regimes and that there is appropriate oversight of those regimes.

In addition, a more specific set of gaps will hold up development of effective AI certification, including limited demand signal in the marketplace (leading to minimal investment in ecosystem development), uncertainty about the readiness of upcoming standards for the challenges of AI, limited transparency by companies in the "ethical AI governance" advisory space about how they are doing their evaluations, and concerns about AI-specific issues that create novel problems beyond the most commonly used tools in certification.

Finally, there is a gap between discussions and tools centered on governance, standards, and policy as they compare to the realities of the technical implementation of AI systems through highly complex, layered technology ecosystems. There is no well-developed reference architecture or policy architecture for AI across ecosystems that can support guidance about who has which responsibilities, and to whom an auditor or certification body should look to verify a given claim about a given implementation.

In this context, the use of the term "architecture" with both "policy" and "reference" refers to a set of documents that describes courses of action, offers guidance and instructions, and delineates specific requirements in order to serve as a communal foundation for the construction of a program or initiative. The goal of architectures is often to create commonality among stakeholders of varying expertise areas in a collaborative project or initiative.

If left unsolved, these gaps will delay or foil efforts to establish effective AI certification. But the gaps identified don't need to stop progress toward certification capabilities for AI. It is possible to develop highly effective initial certification models that take advantage of existing work on governance, management systems, risk models, and other techniques. We can already gain much value from getting started with the tools that are being launched. The raw material is in place in the form of frameworks, very specific audit-ready processes, and governance guidance. To ensure that certification capabilities for AI continue to advance, it is essential that market players harness these tools and take the necessary steps to develop the wider ecosystem.

Recommendations

With these gaps in mind, CWG has established the following recommendations for government, industry participants, civil society, and academia.

Recommendation 1: Government must lead on establishing objectives, resources, and funding

Standards and certifications support societal objectives and derive their direction and grounding from them. They are not necessarily an effective place to debate and decide what those objectives should be. Absent government input on those goals, standards are likely to either overreach in attempting to decide societal objectives, or to leave too much to the choice of individual implementers. Without a consistent articulation of those objectives, conformity assessment bodies (third parties who do the review work needed for an organization to receive certification) won't know what they are looking for. Purchasers won't be able to compare one certification to the next. Assurance systems won't have the necessary significance in order to provide the trust and assurance that they are designed to deliver. Shortcomings in the availability of effective standards and certifications will create uncertainty for developers about where the boundaries are, hampering innovation. The same shortcomings will also impair development of consumer trust in these new technologies, slowing market growth and potentially leading to precipitous actions as that distrust impacts political decisions.

Having developers and implementers choose whatever framework suits them is filled with its own risks. Algorithm Watch has cataloged over 170 different frameworks for "Ethical AI." There is some commonality across these existing frameworks, but there is much that differs as well. The legitimacy and credibility of a system meant to assure that participants meet certain performance standards requires that there is a foundational set of objectives that all parties can rely on. Without the certainty provided by such a foundational set of objectives, AI developers are left to wonder which standards will meet the "right" government or societal objectives—a state of uncertainty that presents a serious impediment to innovation. Though some AI developers may be comfortable taking on the risks associated with designing, developing, and deploying AI systems or products without clear guidance on whether the ethical AI framework they have chosen to follow will be accepted as "correct" or legitimate by government actors, many will not.

Governments and intergovernmental collaborations must step in to provide grounding for AI certification, including investing in internal capabilities and workforce. They are the only participants with the legitimate backing of the democratic process and societal recognition. The certification process for AI will need clear, well-articulated objectives to deliver a consistent, meaningful assessment of AI implementation. Governments can promote both ethical AI and innovation by stepping up to lead on supplying concrete objectives. They can generate important signals to the market by licensing, approving, or providing safe harbour status for market providers that they evaluate to meet government objectives, as suggested in the regulatory markets model.[6] In addition, for government to lead on these fundamental efforts to advance societal objectives, government must also invest in its internal capabilities and workforce in order to house experts who understand conformity assessments, certifications, AI and its audits, development, and oversight.

Recommendation 2: Government and other organizations with large procurement budgets should support market development

Government at all levels can and should play a key role in generating and signaling demand for certification—for example, through regulation and procurement. Driving demand for AI certification will drive innovation in the certification process, regulatory technologies, and trustworthy AI itself. Governments can create market demand for certification by mandating certification of AI systems for high-risk use-cases and by integrating it into public procurement of AI. One of the most important pathways to drive demand for effective certification could be for key government agencies to signal their intent to begin prioritizing (or favouring) certified systems in their procurement processes for AI systems. Even relatively informal notice of a requirement that may not be effective for a year or two would help create momentum if it came from an agency of meaningful size.

Similarly, large companies—even large IT companies—purchase much of their IT from a complex supply chain. Signal to that supply chain would create a meaningful message to the market about demand for AI certification. If governments can signal to industry their plans to pursue certification as at least one facet of their approach to AI governance, industry will almost certainly respond with an outpouring of certification companies and services to meet perceived future demand. Yet even independent of government intervention, stakeholders with significant purchasing power can send clear signals that they—and thus, likely a significant percentage of a given market—will be seeking to procure AI services that have been certified. The motivation for that signal may come from regulators or may come from thoughtful industry discussion (perhaps based on enlightened self-interest among some players) about the meaning of accountability for large organizations and the need for appropriate diligence pushed into their supply chains. Regardless of where this signal comes from, it will provide a strong and necessary catalyst for the advancement of AI certification offerings.

The private sector has recently started to offer another path to develop the market for AI certification: insurance. For example, Munich Re, a leading global provider of reinsurance, offers insurance covering certain risks of AI underperformance. That insurance is supported by a separate AI validation service that is working toward authorization to deliver AI certification against several different measures.[7]

Similarly, Armilla AI has begun providing third-party AI verification and warranty solutions for AI/LLM systems, in partnership with global reinsurers Swiss Re, Greenlight Re, and Chaucer. The warranty acts as a guarantee on key model metrics, for instance, related to performance, fairness, and robustness. Third-party warranties are emerging as a market-based solution for building trust in AI, and a powerful complement to certifications.

One of the most important pathways to drive demand for effective certification could be for key government agencies to signal their intent to begin prioritizing (or favouring) certified systems in their procurement processes for AI systems.

Recommendation 3: All stakeholders need to invest time and resources to get the foundations in place.

Certification will require clear standards against which a given AI system or process can be certified, and a clear process to conduct an audit of whatever is being certified. Though this task is under way, it is by no means complete; there is more to do to reach development of key standards and the adaptation of practices to match the challenges and demands of certifying AI systems. Authoritative AI standards will continue to be developed and evolve for many years to come. In the near term, to build certifications that apply across sectors and scenarios, several things need to happen—all of which take commitment, time, and resources from key stakeholders. Recommendations for setting up foundations include:

1. Complete the development of a viable first set of internationally recognized documents to support certification.

Typically, these initial foundational standards take the form of a wider group of documents that include core principles, definitions, audit schemes, audit standards, risk management methods, and so on. The current work at the Joint Technical Committee of the International Organization for Standardization and the International Electrotechnical Commission (ISO/IEC/JTC1) SC42 to create an AI Management System standard[8] forms an important part of the framework, but significant further work is needed to develop other tools (for example, standardized impact assessments) and enable viable certifications that reach beyond an organization's AI management system. Importantly, it also isn't clear that certification of management systems will be enough to satisfy some of the developing legislative requirements, which may call for certification of specific AI systems (the end product) rather than certification of the management system that helped create the AI system, especially for high-risk AI use cases.

2. Enable joint certifications.

Management systems certifications (under procedures described in ISO/IEC 17021[9]) are often misunderstood as product certifications (under procedures described in ISO/IEC 17065[10]). For AI assessments to be useful in many areas (especially developing regulation), they need to deliver the benefits of both: they must apply to specific products or implementations, but must also take on important attributes from management systems certification. Indeed, as we look at recent, highly visible developments in generative AI, management system and risk assessment techniques will be critically important in enabling certification of products that contain substantial unknowns. Yet the specific context that product certification requires will be imperative to consider given the very fact-specific use cases and risks that are developing. Certification experts need to come together to create joint management systems and product certification rules that will make sense for users, developers, and regulators of AI. The market and certification bodies will need to adopt an effective operational approach to joint AI certification that is recognized and accepted across the industry. That process may also benefit from clarifications in upcoming AI standards and in the surrounding guidance documents and government rules or guidelines that specifically call for joint certifications.

3. Clarify what frameworks can be used for conformity assessment (and what can't).

Standards bodies and others publishing guidance for ethical AI (for example NIST's Risk Management Framework) need to be very clear about whether the specific instrument is intended to be used for any form of conformity assessment. Most are not, but many are already being misused for this purpose. This misuse poses a serious risk, as the 'requirements' that may be set out in a guiding document not meant to be used as the basis of a conformity assessment may be much less thorough than documents created with this intended use in mind. This is a clarification that could be made by the groups publishing the non-normative instruments, or by others in guidance documents.

Recommendation 4: Stakeholders should move quickly to advance the state of the art from these foundations.

4.1: Build transparency and data availability into next generation standards

Assurance methods in AI certification need their own system and methods to confirm that they are achieving their intended objectives. At the same time, private systems that are not validated through professional licensing or accreditation (such as many regulatory technology solutions) will need to develop sufficient transparency about their methods to deliver the needed confidence, credibility, and legitimacy in their results. Innovation in systems to oversee other systems will be important to the development of trusted AI. Regulatory technology holds significant promise to enable AI developers to innovate and compete at necessary scale and speed while staying within appropriate boundaries. At the same time, that regulatory technology itself must pass the "trust test." As a developing new field, the regulatory technology community has not established its own principles for how that trust should develop. Companies we have met with all have acknowledged aspects of this issue, though their proposed mitigations have varied—a clear statement of principle about what level of transparency is appropriate and what purpose it should serve may be an important starting point, especially if some of these companies support that statement and follow through on it.

In addition, visibility into real world implementations of AI systems is critical to understanding the effectiveness of the entire ecosystem of AI standards, certifications, principles, and regulation. Effective AI certification also depends on well-developed research into risks and consequences from various uses and abuses of AI technology. Independent researchers need access to confidential data and AI methods to study risks and harms that are important upstream factors to allow governments to regulate appropriately. Creating a clear expectation in AI standards about the principles for that data access and how it would be implemented would provide significant value to independent research.[11] These expectations do not need deep changes in intellectual property or privacy regimes. They can be implemented through limited exceptions that are built into next generation AI standards and near-term changes in procurement requirements for government and commercial AI systems.

4.2: Government, academia, and industry should develop a reference architecture to serve AI policy and certification

As noted above, there is no well-developed reference architecture for AI across ecosystems that can support guidance about who has which responsibilities, and to whom an auditor or certification body should look to verify a given claim about a given implementation. A pair of key documents in organizing the market for cloud services, including roles in contracting, certification, and security operations, were created by the US's National Institutes of Standards and Technology (NIST) in 2011. The definitions and reference architecture in these documents established grounding for the layers and types of cloud service and the roles that different participants would play.[12] That reference architecture was not a detailed technical specification, but a notional model that allowed players to identify roles and responsibilities, as well as risks, in a practical, consistent way.[13]

We need a similar reference architecture—or at least a regulatory and assurance architecture—for AI. Indeed, Microsoft has recently proposed the idea of a "regulatory architecture" in a white paper identifying a layered model for policy.[14] We believe a clear reference architecture for the rules of the game is necessary beyond regulation—for policy, for accountability mechanisms, for insurance, and even for allocation of responsibilities in contracts and ultimately in the courts in determining liability. The appropriate government entities should proactively play a convening role to develop a broader, vendor-neutral model that can serve for purposes of assurance, regulation, and roles and

A clear reference architecture for the rules of the game is necessary beyond regulation—for policy, for accountability mechanisms, for insurance, and even for allocation of responsibilities in contracts and ultimately in the courts in determining liability.

responsibilities.

4.3: Government, academia, and industry must develop a focused research agenda to advance AI verification and validation tools for certification

The first iterations of widely-used AI certification will undoubtedly be incomplete, even though they also stand to raise the bar substantially on the responsible development and use of AI. AI raises a range of issues that require significant changes to thinking on measurement, testing, verification, validation, and how we conclude that any particular attribute is true and repeatable for a given system.

The stochastic nature of AI and the often opaque nature of the data that trains it have created a truly confounding set of math, measurement, and social impact problems that affect all forms of audit and verification. **This requires research into key sociotechnical topics about AI's mismatch with the discipline of certification and product conformity assessment that has developed over the past century.** Some examples discussed by the CWG of the ways in which AI's characteristics can evade established certification and conformity assessment processes include:

- Versioning and release management in software development means that AI systems change gradually over time; can past iterations of ML models be reproduced and what does this mean for certification, which might focus on static or discrete software?[15]
- Large-scale data (too large for traditional data-processing software) or free-range data (inconsistent in formats, often resulting from unstandardized data entry) often cannot be tested, documented, or managed well.[16]
- Evolving datasets under changing conditions (including feedback loops by users, adversarial attacks, etc.) can cause AI systems to degrade or diverge from intended behaviour over time.[17]
- Pre-trained opaque algorithms available "off the shelf" pose a problem in their lack of connection between development context and implementation context (and these are often themselves trained on large-scale, free-range data.)[18]
- AI models can produce predictive rather than deterministic outcomes; the former includes uncertainty and probabilistic results, which pose a problem for certification insofar as they are not definitive.[19]
- Decision-making power vested in AI systems raises important questions about public trust and the social roles granted to automated systems.[20]
- Learning systems deliver results that may differ from expectations—sometimes due to poor calibration of the reliability or accuracy of a system.[21]
- Incomprehensible math (or non-math processing) may make decisions we cannot understand or predict.[22]
- Composed systems with opaque building blocks in an extended supply chain of AI systems can result in failure to consider systemic behaviour.[23]

In the near term, certification efforts will need to work around the above (and related) issues through process analysis, impact assessment, risk treatment, and ongoing surveillance of systems. There are important gains from putting workarounds in place that would allow certification to develop while we find better techniques to resolve some or all of these issues. But in the long term, these issues deserve in-depth research, with a very specific intent to develop results that will help with governance, oversight, and improvement of the behaviour of AI systems.

Government should also play a crucial role in driving the research necessary to address AI-specific attributes of conformity assessment, both through direct funding and research efforts at government institutes, and through programs, rules, and incentives that aid qualified researchers in gaining access to the kind of information needed to study risks and harms of AI to better inform societal requirements.

Conclusion

Our relationship with AI is in its infancy. Recent advancements in applications using AI for text and image generation have given us new visibility into some of what may come. The opportunity to advance societal goals and economic objectives through well-developed emerging AI technologies is important—we must not let it pass us by. But neither should we let the current window close on us before getting thoughtful AI certification systems in place to foster innovation and create guardrails—and to keep humans at the centre. As an interdisciplinary, multinational group of experts, CWG feels a responsibility to articulate this opportunity, flag notable gaps, and make key recommendations to accelerate AI certification and unlock the value it can bring at this important inflection point. The real work here, and the real chance to make a difference, rests with the key stakeholders: government, industry, academia, and civil society. We offer our support for this work, and our time and effort where it is appropriate to call upon us to help.

Appendix 1: Definitions

Assurance ecosystems, or ecosystems of trust, are made up of several different components and are intended to provide consumers with justified trust in a particular product or service. For example, the UK has laid out a roadmap to an effective AI assurance ecosystem, detailing the need for third-party auditors, certification, assessments, and regulation to create a balanced ecosystem in which consumers can trust that any AI systems in use have met a certain safety threshold. Although proposals for assurance ecosystems may differ, standards are a vital component of any such ecosystem.

Conformity assessments confirm whether a service, system, or product adheres to the requirements of a particular standard or regulation. Such requirements may include, for example, performance, safety, efficiency, effectiveness, reliability, durability, or environment impacts.

Certification is a process through which an independent body attests that an organization or its personnel, systems, or products meet objective standards of quality or performance, typically through the issuance of a "mark" or "label."

Impact assessments evaluate the impact a particular activity or system could have. For example, an impact assessment of an AI system that decides who will receive a loan might identify whether and how much those seeking loans might be affected. Impact assessments for AI may build off existing impact-assessment frameworks in fields such as environmental protection, human rights, or data protection.

Standards are documents that set out established practices arrived at by consensus and approved by a recognized body. They provide for common and repeated use, rules, guidelines, or characteristics for activities or their results, and are aimed at achievement of the optimum degree of order in a given context. Standards are typically voluntary but can become mandatory when enforced by laws or regulations—for example, for health or safety reasons.

Notes

[1] G7 Hiroshima Leaders' Communiqué, 20 May 2023 (whitehouse.gov).

[2] Legislation and regulation will play a key role in the demand for certification, principally because legislators and regulators recognize they need to take advantage of certification as a trust and monitoring mechanism across the marketplace. The EU AI Act reached a key milestone in December 2023, as the Council presidency and the European Parliament's negotiators reached a provisional agreement on the proposed Act, and in March 2024 European Parliament voted to approve the Act, paving the way for its implementation in the coming months. At the time of this paper, Canada's Artificial Intelligence and Data Act is at consideration in committee at the House of Commons. Action has also started on bi-partisan legislation in the US Congress, as the Federal Trade Commission and other agencies look at ways to reduce automated discrimination within their remits. The UK, Brazil, Canada, and South Africa are all taking steps toward specific provisions regulating decision-making algorithms. Many of these regulatory and policy approaches lean heavily on conformity assessment and certification as a foundation for trustworthiness in AI and a way to provide some level of independent evaluation that does not require ex ante reviews that take up scarce regulatory resources. They will drive the need for certification in the market; a key question is whether the market will be ready when the rules demand certification.

[3] IEC 61508-1:2010: Functional safety of electrical/electronic/programmable electronic safety-related systems - Part 1: General requirements (iec.ch).

[4] See for example: ISO/IEC DIS 42001: Information technology - artificial intelligence management system (iso.org).

[5] See for example: Robert Fish, "Can Ethics be Standardized? Creating Modern Standards for Ethical Autonomous and Intelligent Systems," 15 March 2019, IEEE Communications Standards Magazine (standards.ieee.org).

[6] Gillian Hadfield and Jack Clark, "Regulatory Markets: The Future of AI Governance," 11 April 2023 (arxiv.org).

[7] See the "Insure AI" product (munichre.com).

[8] Preview: ISO/IEC DIS 42001: Information technology - Artificial intelligence management system (iso.org).

[9] See the standard series beginning at: ISO/IEC 17021-1:2015: Conformity assessment - Requirements for bodies providing audit and certification of management systems - Part 1: Requirements (iso.org).

[10] ISO/IEC 17065:2012: Conformity assessment - Requirements for bodies certifying products, processes and services (iso.org).

[11] We recognize that to reach some of the most important issues this data access requires navigating existing rules about private data, or obtaining an exception to those rules. We believe that including the expectation in developing AI standards will help move this ability forward in a way that is safe for data subjects and advances the ability for researchers to play a key role in investigating risks and harms in AI systems.

[12] National Institutes of Standards and Technology (NIST) Special Publication #800-145, "The NIST Definition of Cloud Computing" (nist.gov).

[13] As examples, large cloud service providers like Amazon Web Services (AWS) and Microsoft Azure have built on the NIST framework to create a "shared responsibility model" to help communicate, in a highly simplified manner, roles between cloud providers and large customers in managing cybersecurity responsibilities.

[14] "Governing AI: A Blueprint for the Future" (microsoft.com)

[15] Note that the United States Food and Drug Administration has recently released recommendations for this critical area for medical device software: "Marketing Submission

Recommendations for a Predetermined Change Control Plan for Artificial Intelligence/Machine Learning (AI/ML)-Enabled Device Software Functions" (fda.gov).

[16] As a short summary, see Jennifer Bryant, "Generative AI: A 'new frontier'" (iapp.org).

[17] See for example: Rohan Taori and Tatsunori B. Hashimoto, [2209.03942] Data Feedback Loops: Model-driven Amplification of Dataset Biases (arxiv.org).

[18] For an introduction to work on this set of issues, see: "Introducing the Center for Research on Foundation Models (CRFM)" (stanford.edu).

[19] Donald Firesmith, "The Challenges of Testing in a Non-Deterministic World," on the blog of the Software Engineering Institute at Carnegie Mellon University (cmu.edu).

[20] Theo Araujo, Natali Helberger, Sanne Kruikemeier & Claes H. de Vreese, "In AI we trust? Perceptions about automated decision-making by artificial intelligence" (springer.com)

[21] Boris Babic, I. Glenn Cohen, Theodoros Evgeniou, & Sara Gerke, "When Machine Learning Goes Off the Rails," in Harvard Business Review Magazine, January-February 2021 (hbr.org)

[22] See, for example: Will Knight, "The Dark Secret at the Heart of AI," in MIT Technology Review, 11 April 2017 (technologyreview.com).

[23] Casey Clifton, Richard Blythman & Kartika Tulusan, "Is Decentralized AI Safer?" (arxiv.org)

About

About the Schwartz Reisman Institute for Technology and Society

Located at the University of Toronto, the Schwartz Reisman Institute for Technology and Society's (SRI) mission is to deepen knowledge of technologies, societies, and what it means to be human by integrating research across traditional boundaries and building human-centred solutions that really make a difference. The integrative research SRI conducts rethinks technology's role in society, the contemporary needs of human communities, and the systems that govern them. SRI is investigating how best to align technology with human values and deploy it accordingly. The human-centred solutions SRI builds are actionable and practical, highlighting the potential of emerging technologies to serve the public good while protecting citizens and societies from their misuse. SRI's mission is to make sure powerful technologies truly make the world a better place—for everyone.

About the Responsible AI Institute

The Responsible AI Institute (RAI Institute) is an independent, not-for-profit organization focused on developing and scaling responsible AI programs within organizations. As a member-driven and community-focused entity, the RAI Institute convenes and works closely with industry, academia, policymakers, and regulators to enhance responsible AI practices. The RAI Institute also offers a range of services including maturity assessments, system-level assessments, supplier assessments, policy and governance support, and training.

Since its establishment in 2016, the RAI Institute has led the field in advancing an AI certification program at the AI system level. The RAI Institute also maintains an AI Regulatory Tracker, convenes AI working groups on Sustainability, Automated Employment, and Financial Services, and publishes guidebooks on topics such as enterprise AI governance, deepfakes, and comparing commercially-available LLM providers.

About the World Economic Forum's Centre for the Fourth Industrial Revolution

The Centre for the Fourth Industrial Revolution (C4IR) is a platform for multistakeholder collaboration, bringing together public and private sectors to maximize technological benefits to society while minimizing the risks. The World Economic Forum launched the first C4IR in San Francisco in 2017, shortly followed by centres in Japan and India. The network now includes centres in Austin, TX (Centre for Trustworthy Technology), Azerbaijan, Brazil, Colombia, Detroit, MI (US Centre for Advanced Manufacturing), Germany (Global Government Technology Centre Berlin), Israel, Kazakhstan, Malaysia, Qatar, Rwanda, Saudi Arabia, Serbia, Telangana (India), and the United Arab Emirates.

ARTIFICIAL INTELLIGENCE CERTIFICATION

Unlocking the power of AI through innovation and trust

A report from the Certification Working Group (CWG), a collaboration established by the Schwartz Reisman Institute for Technology and Society at the University of Toronto, the Responsible AI Institute, and the World Economic Forum.

April 2, 2024

srinstitute.utoronto.ca
hello@torontosri.ca

@TorontoSRI