Responsible Artificial Intelligence (RAI) Certification Beta

April, 2021

Introducing the RAI Certification Beta

With all promising innovations, we are often intrigued by the potential they have to revolutionize our lives and society while simultaneously maintaining a healthy skepticism because they are different, creating a significant amount of change, or maybe just seem too good to be true.

Artificial Intelligence (AI) is in a class with technologies that have reshaped our society like the steam engine, telecommunications, and the Internet. And while these tools have demonstrated the capacity to advance life as we know it, even decades later we are living with their consequences, intended or unintended, having devastating impacts on our environment, privacy, and democracy.

While many consider AI a new technology, the concept of Artificial Intelligence dates back to 1956 when John Mcarthy hosted the first conference on the subject. Since then, there has been a long history of both real and imagined horrors that could occur due to the advancement of AI.

With AI quickly becoming an integral part of our daily lives (with varying degrees of visibility) from the music we are recommended, the predictive text we write, the information we see and are suggested to share, the perfect heating in our homes and businesses, the services and credit we can access, and determinations on whether or not we get hired, AI is there.

While we whole-heartedly believe in the numerous benefits of these systems and many more not listed, we have been keeping track of cases where AI has gone wrong. As such, we are part of a burgeoning community of policy makers, technologists, engineers, investors, researchers, and business leaders who are raising questions about what type of oversight is needed to ensure that with the advancement of these technologies, human rights are respected, individuals and organizations are safe, and the planet is better-off not worse-off.

The Organization for Economic Co-operation and Development (OECD) has developed a comprehensive collection of these calls to action through their AI policy observatory. While many governments and regulators have started to take stronger stands on the oversight of AI technologies, the most notable is the European Union's (EU) recent Proposal for a regulation laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act). The EU is not alone, in the same week as this proposal was released, the US Government's Federal Trade Commission, warned in a blog post "Aiming for truth, fairness, and equity in your company's use of AI", Hold yourself accountable – or be ready for the FTC to do it for you.

In response and alignment with these demands, the Responsible AI Institute (RAI) in partnership with the World Economic Forum (WEF)'s Global AI Action Alliance and the Schwartz Reisman Institute for Technology and Society at the University of Toronto (SRI) having been leading the development of a community-driven, measurable, and independent certification mark for the responsible and trusted use of AI systems.

This document provides an overview of our work to date and the current approach of the Responsible AI (RAI) Certification Beta. As well, it highlights work being done by the community, and our partner organizations, to enhance and advance these efforts. And most importantly, outlines how individuals and organizations alike can get involved to co-build and validate this work.

If you are interested in learning more or joining this important work, please contact us at admin@responsible.ai.

Our objective

We believe that for the most part, those who build, acquire, and use AI systems do so with the best intentions in mind. However, even good intentions can lead to serious consequences. Whether we are prioritizing speed of innovation over ethics, directly responding to market and customer demands without thinking about the global ramifications, or lack clear guidance, regulations and standards to follow, we are all susceptible to forging ahead and applying emerging research and science with a limited understanding of the potential consequences of AI systems.

We don't want to make AI oversight guesswork. We think it should be as simple and straightforward as possible. This is why we are dedicated to building a comprehensive and independent certification program that is grounded in human rights respecting principles. Ensuring that it is practical and measurable, is internationally recognized, and is built with trust and transparency.

Lastly, as noble as we believe many individuals and organizations to be. We all question those who mark their own homework. If increased trust and transparency is what the public are advocating for, then an independent certification is the way to address these concerns.

We know that these are lofty objectives, but we have a concrete plan to make this possible. Most importantly is recognizing that AI is everywhere, and it means a lot of different things. This is why, we don't attempt to define it too narrowly. We think that there should be a responsible use of all technology, but especially those technologies which have the capacity to adapt and learn. This is why we have decided to start with some key use cases which have high prevalence and the most considerable concern for individuals and society. For the time being, this includes, looking at AI systems in the financial, health, and labor sectors.

How it work and why it's needed

Build

A comprehensive community validated framework

Accredit

Framework assesed by National accreditation bodies

Audit

Audits performed by trained and independent auditors

Certify

Four levels of RAI Certification are awarded

RAI Certification Framework dimensions

- Accountability
- Bias and Fairness
- Data quality
- Explainability and Interpretability
- Robustness

Avoiding bias and discrimintation

The aforementioned calls for oversight are based on significant research that has demonstrated various different AI systems to be biased and discriminate against certain demographics. From what we've seen so far, it is typically demographics who suffer from historic or systemic discrimintation. Our concern is that these systems are poised to expedite and ultimately exasperate these issues. Working with governments, standards organizations, international bodies, researchers and industry leaders the RAI Certification adheres to emerging regulations, standards, and best practices being researched and implemented.

Consistent compliance with regulations

As we've seen with rules and regulations throughout history, not all of them are clear. In large part, this is why we have an entire judicial system to help us interpret what the intent of these rules are. Emerging regulations and standards, like the EU's proposed Artificial Intelligence Act which is already being likened to the GDPR for AI makes a lot of important and necessary declarations, however, also leaves a lot of questions for what consistent implementation would look like. Especially given the broad scope of AI systems. The RAI Certification establishes a consistent way to demonstrate compliance with these emerging regulations and standards.

Prevent brand and reputation damage

As mentioned above we believe very few organizations have the intent of actively harming any of their clients, customers, or users. As we've seen with facial recognition, automated hiring, autonomous driving, and more, the intent does not always match the outcome. By taking a comprehensive approach, rooted in design-thinking practices, the RAI Certification helps to identify potential gaps in governance, oversight, and performance of AI systems. Additionally, it raises key discussion points for the whole organization involved in the design, development, and deployment of AI systems.

Build customer trust and loyalty

Trust can be earned in a variety of ways, over time, through a strong reputation, and in many circumstances because an endorsement comes from a reputable source. By establishing a comprehensive and authoritative certification built and validated by experts in the community, accredited by National accreditation bodies, and delivered by independent auditors, the RAI Certification provides your customers with the confidence they need to know that your team took all the necessary steps to mitigate unintended consequences and harm.

Build higher performance data and AI systems

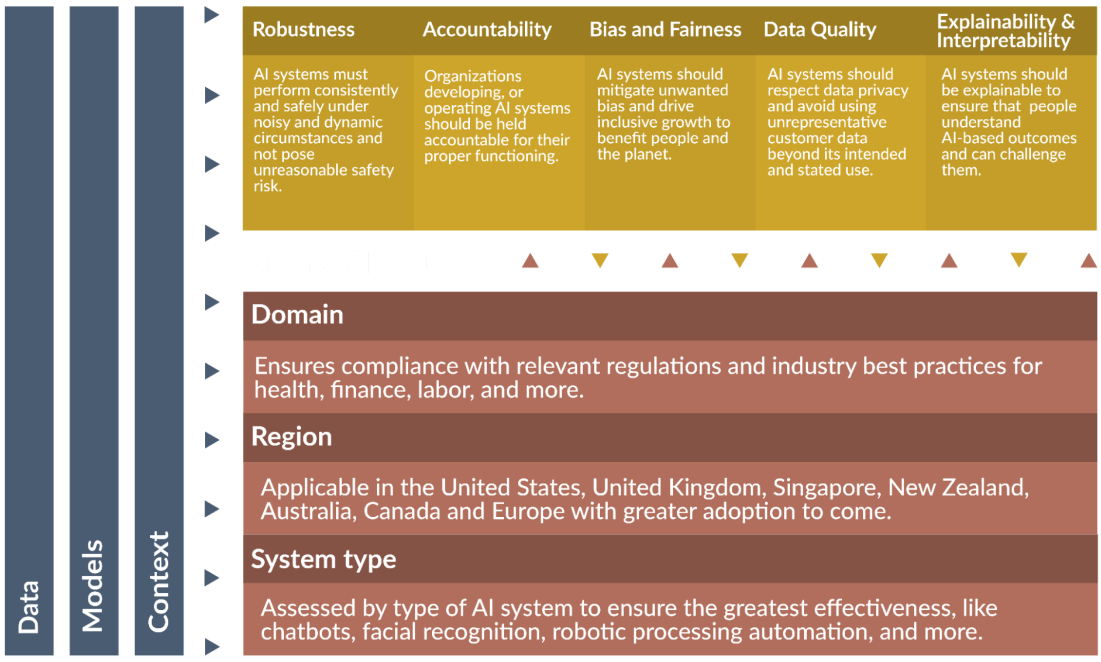
By creating a measurable and repeatable assessment for AI systems, the RAI Certification takes the guesswork out what good looks like. Bringing together the standards, regulations, and emerging best practices into one comprehensive assessment framework the RAI Certification aims to make it as easy as possible for practitioners to do what they are good at and advancing their tools and technologies without worry.

Scope

Following extensive research and discussions with those who are building and using AI systems, we got to the heart of what needs to, and can be certified. While it's possible to certify individuals (which we highly endorse) and organizations can be ranked, we think it's of ultimate importance to review the system based on when, how, and to who they are being deployed to. This is why the RAI Certification assesses the data, model, and contextual deployment of the system as these are all factors in the efficacy, fairness, or usefulness of the system.

Incredible and expansive research has been done to identify various different principles, practices, and concepts which should be followed to ensure the responsibility of these systems. We did not want to add to these growing body of research, as such, we have been inspired by all of them, but remain grounded in OECD's AI principles. These systems are assessed against five key dimensions, robustness, accountability, bias and fairness, data quality, and explainability and interpertability.

Finally, we recognize that there are going to be differences in how these systems are used based on their domain, region, and technology or system type. While we can't begin to capture all of these nuances to start, we do hope to expand our understanding and assessment through key pilot projects to ensure the RAI Certification works for as many systems as possible. Below we outline the key use cases we are starting with in the beta phase of this initiative.



Key Audiences

Audience	RAI Certification benefits
Senior Executives/ Executive Review Boards	● Assurance that the AI systems their organization are producing or using are trustworthy, compliant with existing regulations. ● Confidence their teams have taken all necessary precautions to mitigate bias and unintended consequences.
Compliance Officers	● Consistent reporting and information sharing between first and second line of defence within a company.
Procurement Officers	● Ability to compare like-systems. ● Confidence that they are procuring trustworthy AI systems that won't cause the organization concerns when in use.
Regulators	● To ensure compliance with their established regulations
Investors	● Confidence they're investing in AI systems that won't cause harm to people or the planet
Consumers	● Confidence AI systems they are using are protecting their rights, privacy, and civil liberties.
Trusted Integrators	● Key outputs and validation of selected measures for combined reporting

Table 1

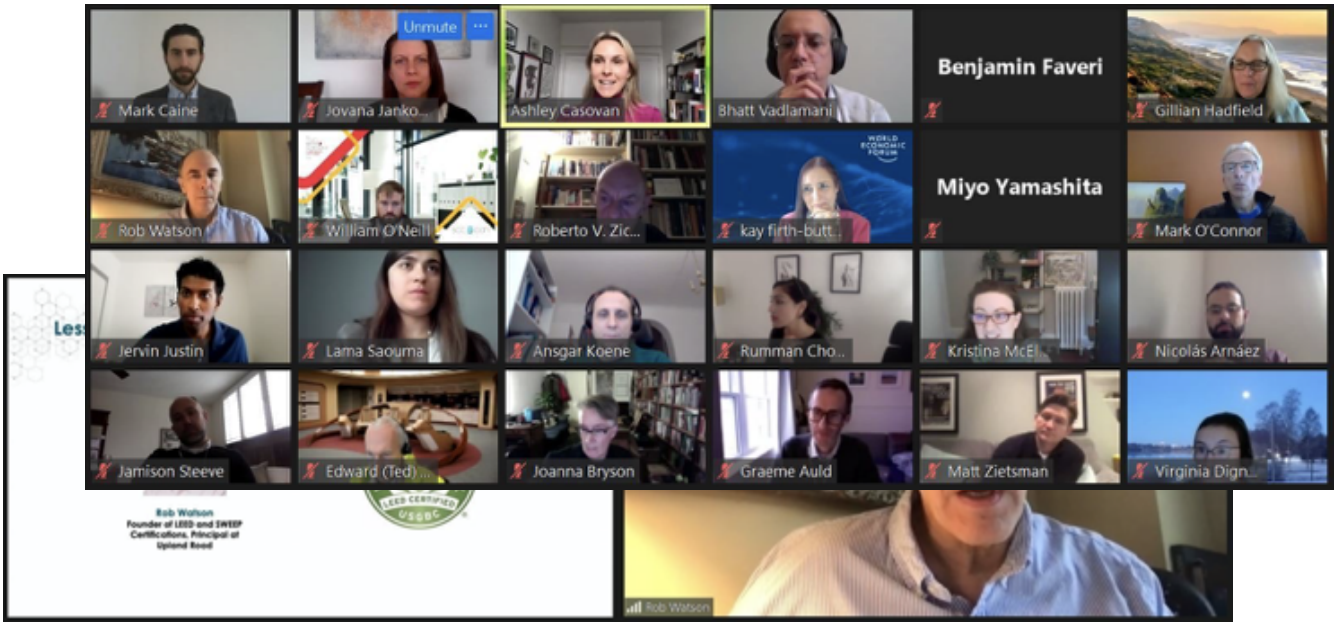
Maintenance

As AI systems have the ability to quickly adapt and learn on their own, one of the key considerations we have identified to test as part of the beta phase is how long the RAI Certification would be valid for. While we hope it will be based on a degree of fluctuation or drift in the system, as opposed to duration of time, it is difficult to know at this point what will be feasible. What we do know is that strong continuous monitoring of AI systems is an imperative.

Work done to date

RAI has been working on building a strong foundation for a Responsible AI Certification Mark Program since the fall of 2019.

Recognizing that a project of this magnitude needs to be built by the community for the benefit of the community, we launched the RAI Certification Working Group December 2020 with WEF and SRI. Following the kick-off, we had an outpouring of support and interest in this work which led to numerous discussions with technical, data, governance, ethics, and industry experts.



In advance of this workshop our research started with the below investigations. In many of these circumstances, we realized that our research would be helpful to share with the community, as such, continue to be publicly available and are maintained based on our interest and community interest:

- Where AI has Gone Wrong map and dataset
- Responsible AI Documents visual and dataset
- Responsible AI Landscape Review
- AI Standards Review

This research informed a set of assessment questions, at the time referred to as the Responsible AI Trust Index, these questions were the starting point for a certification mark program. Additional documents speak to the development of domain and regional specific assessment.

A publicly available open source version of the Responsible AI Trust Index and the Responsible AI Design Assistant are already available through the RAI website. Ongoing review and maturation of this tool is underway to ensure it supports organisations who are interested in implementing responsible AI practices. In the future, using the Responsible AI Design Assistant will set them up for success for the Responsible AI Certification Mark.

Additional resources can be found here:

- Responsible AI Design Assistant Tool for organizations to use early and often to prepare themselves for certification.
- Responsible AI Trust Index Blog Post
- DRAFT: Responsible AI Trust Index Questions
- Business plan from a trusted International Accreditation organization (Standards Council of Canada who is able to accredit certification organisations world-wide).
- Responsible AI Independent Review Guidelines

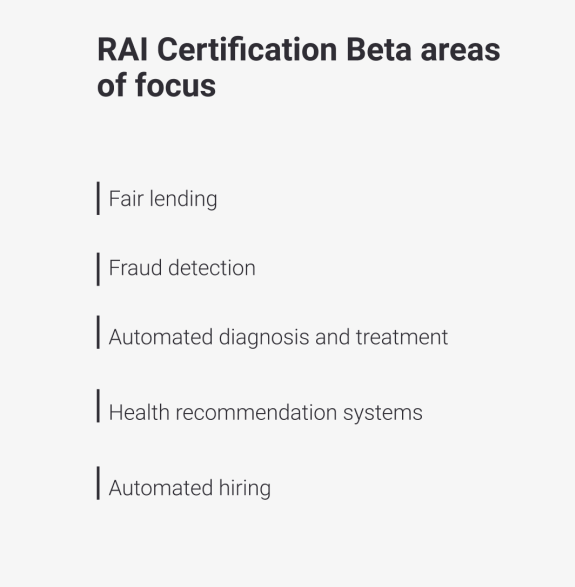
Moving forward

Areas of focus

We are doing this work for the community and with the community. The current assessment is comprehensive, but subjective, and process driven. The objective is to develop a more objective assurance style assessment. In order to do so, we have to start, by selecting key areas of focus based on:

- Issues most critical to society (eg. mitigating discrimination to vulnerable populations)
- Highest uses of AI systems in industry (high demand)
- Current regulations which could apply
- Anticipated regulations

As such, we've decided on the following use cases for the RAI Certification beta: Fair Lending, Fraud Detection, Automated Diagnosis and Treatment, and Automated Hiring.



Futures areas of focus

This will continue to evolve as we see trends in arising challenges.

Retail

- Behavioural nudging

- Targeting
- Privacy
- Lack of autonomous decision making

Finance

- Fair lending
- Accuracy in lending and services
- Fraud detection
- Privacy

Health

- Access to service (eg. Accuracy in provider matching and pricing)
- Accuracy in diagnostic and treatment (eg. Personalized medicine)
- Fraud detection
- Back-office efficiency (eg. Management of electronic medical records)
- Privacy

Social Services

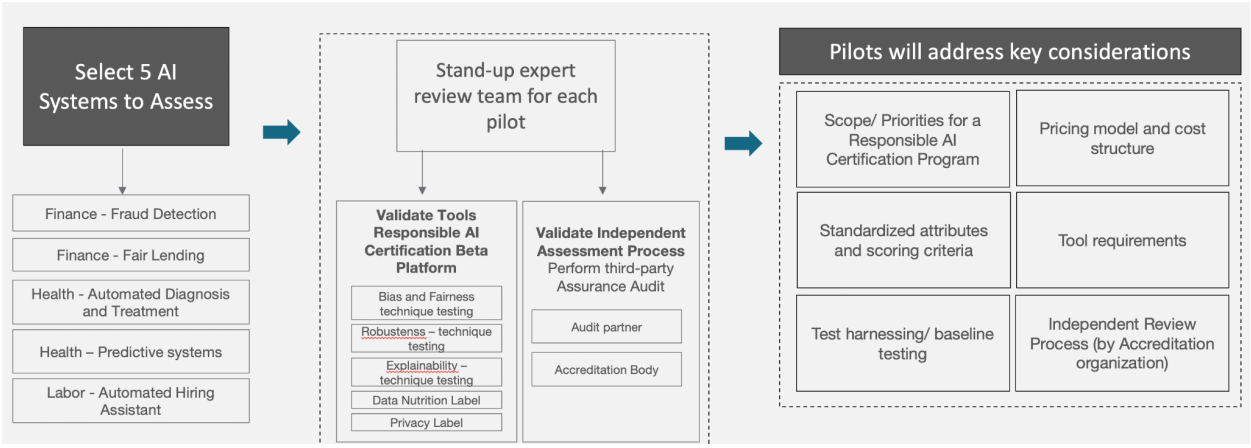
- Biased predictions
- Social Media
- Disinformation
- Privacy

Social Media

- Disinformation

Pilot approach

Dedicated to remaining independent and authoritative, the RAI certification will be delivered by third-party auditors. As such, as part of the RAI Certification beta, the framework will undergo a conformity assessment performed by National Accreditation bodies.



The insight gained from each of these pilots will allow us to better understand the scope of our efforts.

Badging

As noted above, while the RAI certification will be focused on the delivery of a comprehensive certification of AI systems, as you can imagine, verification of these efforts will be dependent on several sub-components including existing compliance with standards and protocols as well as proof of education and training. As consistent education standards both at the university and professional level are still under development, part of the RAI Certification Beta will be to partner with organizations delivering this type of training.

We intend to recognise and co-develop badging, particularly for verification of education and training associated with different aspects of the RAI Certification assessment. One of our key partners for this will be Equal AI, who are in the process of establishing a Responsible AI badging program for Executives.

More information on this to come soon.

Reporting

We see reporting being a significant need for all of the audiences listed above in Table 1. Through initial pilot testing, we have started to identify the types of reports which would be required by each audience. Including what information is needed, to what degree of detail, and for how long will it be valid.

This is a sample report based on initial beta testing performed by our members and community leaders.

As with the whole process, we will continue to work through each of the working groups to validate the output of the RAI Certification, including the reporting mechanisms.

An essential component of these reports will be data labels produced by our partners at the Data Nutrition Project. Privacy validation powered by BEACON Trust Network, and model reviews completed by model validation organizations including CognitiveScale and Arthur.AI. Finally, incorporating completion of badging discussed above.

Next Steps

We need your expertise and support! We are in the process of identifying pilot organisations and experts for our five working groups based on our key areas of focus.

For pilot organisations we would like you to:

1. Select a use case
2. An independent audit team will be selected based on use case
3. Audit team will use Certification Beta Platform to evaluate the use case
4. This assessment will produce a score card, data labels, and a privacy label

Pilot organisation will either:

- Complete the assessment on their own OR
- Provide evidence required by the assessment to the independent auditor

Evidence includes:

- Documentation about the development of the system (the requirements)
- Results of bias, explaniability test (evaluation system will be provided as part of the assessment platform)
- Sample data used by the system for independent evaluation (alternatively an attested interview with Data Nutrition Project can be done)

Determining Sources of funding

In their paper, "Towards Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims, Gillian Hadfield et al[1] recommend that "A coalition of stakeholders should create a task force to research options for conducting and funding third party auditing of AI systems."

Resources are required to build a certification mark program, both to do the initial research, but also to ensure adoption and ongoing administration. While organisations like the US Green Building Council have grown organically over a long period of time, AI systems are advancing extremely quickly. There is a need for oversight to compliment public governance mechanisms as soon as possible.

One of the key tasks of the working group will be to determine mechanisms for funding the development of the initial certification mark program and a plan for the ongoing licensing or other approaches to provide ongoing sustainment of the program.

[1] Miles Brundage et al. (April 2020). *Toward Trustworthy AI Development. Mechanisms for supporting verifiable claims* <https://arxiv.org/abs/2004.07213>

Resources are required for:

- Management and Operations
    - Coordination
    - Oversight of projects
    - Management of working group
- Research and Analysis
    - Research on existing and future frameworks
    - Integrate feedback from consultations and working group participation
- Hosting
    - In person consultations and meetings may be required
    - Virtual hosting

Our Partners

RAI has been supported by a diverse group of subject matter experts to advance this work. We can not thank our individual contributors enough for their efforts. To respect their privacy, we have decided not to enumerate each individual, however, the following is a list of the organizations which have helped us to develop and advance this work to date:

- Algora Labs
- AltaML
- Anthem
- American Express
- Bureau of Labour Statistics
- CognitiveScale
- CIO Strategy Council
- Data Nutrition Project
- Deloitte
- EY
- GovLab
- Hyper Giant
- IBM
- InfoTech Research Group
- Jackson National
- Lucid AI
- MILA
- Montreal AI Ethics Institute
- New York State
- Office of AI, Government of UK

- Oproma
- Oxford Internet Institute
- Prudence AI
- United Nations Office for Disarmament Affairs, Pan-Asia
- University of McGill
- University of Texas
- Smith School of Business, Queen’s University
- Standards Council of Canada
- Schwartz Reisman Institute, University of Toronto
- Tilburg Institute for Law, Technology and Society
- World Economic Forum
- Yum! Brands