



Responsible
Artificial Intelligence
Institute

Advancing Trusted AI

October 2022

The Responsible AI Certification Program

White Paper



**Responsible Artificial Intelligence Institute (RAII) - Responsible AI
Certification Program - White Paper**

October 2022

© 2022 Responsible Artificial Intelligence Institute. All Rights Reserved.

Table of Contents

05 About RAI

- 05 Background
- 07 The Benefits of the Certification
- 08 Areas of Focus
- 09 Case Study: Financial Institution

10 Context

- 10 Demand for a Global AI Certification Program
- 12 Laws, Regulations, and Enforcement
- 14 Globally Adopted AI Principles and Frameworks
- 17 Practitioner and Research Insights

18 RAI Certification Program

- 18 Scope
- 20 Delivery
- 22 Dimensions and Subdimensions

26 RAI Community

- 26 Ecosystem
- 28 Our Team

30 References



About RAI

October 2022

Background

The Responsible AI Institute (RAI) is developing one of the world's first responsible AI certification programs. The RAI Certification Program is aligned with emerging global AI laws and regulations, internationally agreed-upon AI principles, research, emerging best practices, and human rights frameworks. RAI is an independent and community-driven non-profit organization building tangible governance tools for trustworthy, safe, and fair AI.

As a member of the World Economic Forum's Global AI Action Alliance (GAIA), RAI joins hands with over 100 government entities, civil society organizations, private companies, and academic institutions to identify and implement tools and best practices that promote responsible AI (World Economic Forum [WEF], 2021). As AI systems are becoming increasingly prevalent, governments, companies, and civil society organizations are grappling with approaches to govern AI systems in a consistent manner. Recent research has suggested that certification programs for AI could serve as an important complement to laws and regulations (Cihon, 2019; Cihon et al., 2020).

Organizations around the world have put forward responsible AI principles (BSI, 2021; Council of



A handwritten signature in blue ink that reads "Ashley Casovan".

Ashley Casovan, RAI's Executive Director, previously led the development of Canada's Directive on Automated Decision-Making Systems, a pioneering policy instrument that set the standard for acceptable government use of AI systems. In addition to her experience, RAI's Certification Program is based upon over three years of research, integration, testing, and lessons learned from RAI members' responsible AI initiatives.

Europe, 2020; European Commission, 2019; ICO, 2021; NIST, 2022; OCC, 2021; WEF, 2020). Accordingly, a general, international consensus on what constitutes responsible AI has emerged. The RAI Certification Program takes the guesswork out of what it means to be responsible, by translating globally adopted principles, standards, and regulations into clear implementation requirements.

The RAI Certification Program is based on a maturity assessment that evaluates AI systems. Recognizing that not all AI systems are the same, this program tailors its tests to specific industries and functions. RAI's initial focus industries and functions are: finance, health care, HR, and procurement.

Informed by those researching, designing, building, deploying, using, and overseeing AI, the RAI team has aggregated extensive information from various perspectives to understand:

- > What responsible AI is;
- > Why we need responsible AI; and
- > How certification can support responsible AI adoption.

RAI's Corporate Members



The Benefits of the Certification

October 2022

RAI's certification benefits different stakeholder within the AI ecosystem:

Key Audiences	Value of RAI Certification for Audience Group
All Stakeholders	Sets a clear bar for global best practices to implement AI responsibly, providing certainty, direction, and actionable next steps.
Senior Executives & Executive Review Boards	Gives confidence that the products and services they are deploying are fit for purpose, legally compliant, of an appropriate quality, and scalable.
Compliance Officers	Enables involvement at the design and development phases, thereby avoiding costly and difficult compliance decisions later in the AI system lifecycle.
Procurement Officers	Provides processes to procure trustworthy AI systems, enabling an organization to deliver quality AI products and services while reducing liability and risk.
Regulators	Enables compliance with established regulations and alignment with proposed regulatory approaches.
Investors	Provides assurance that AI systems are built on recognized global best practices.
Consumers	Gives comfort that rights, privacy, and civil liberties are protected.



Areas of Focus

Financial Services



AI use in financial services is projected to increase exponentially over the coming years. McKinsey estimates that AI systems could eventually deliver \$1 trillion of annual market value. RAI has worked with financial institutions, researchers, and standard setters to develop the RAI Certification Program for Automated Lending Systems (see Case Study below).

Health Care



The AI in health care market, worth \$6.7 billion, is growing rapidly (Grand View Research, 2022). RAI is currently developing certifications for two health care use cases:

- 1) automated pre-authorization for health insurance; and
- 2) applying computer vision to diagnose skin disease.

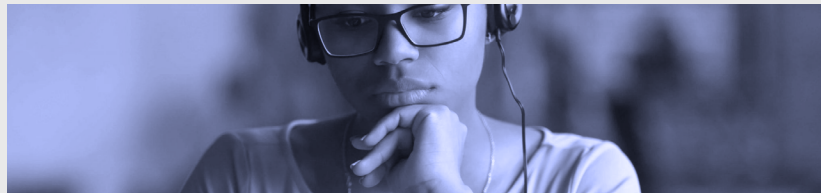
In each of these use cases, RAI is working with industry leaders and researchers at the intersection of AI and health care.

Procurement



Companies are developing frameworks to test and verify the AI tools they procure. RAI is developing a certification program for AI procurement, building upon its pioneering work with the World Economic Forum's Procurement in a Box initiative, its AI procurement pilot with the U.S. Department of Defense's Joint AI Center, and its industry engagements with companies in financial services, health care, and other industries.

Human Resources



Today, companies can choose from over 250 different commercial AI tools, which cover all phases of the HR lifecycle (WEF, 2021). Regulators at all levels - from the New York City council to the European Union - are paying attention (European Commission, 2021; Lee & Lai, 2020; New York City Council, 2021; Marcia & Desouza, 2021). RAI is developing a certification for HR systems, informed by RAI's Working Group on AI in Human Resources, industry engagements, and leading researchers.

Case Study: Financial Institution

Over the past year, RAI worked with a Financial Institution (FI) to calibrate the RAI Implementation Framework to the automated lending use case and to build capacity at the FI in alignment with its responsible AI principles and commitments:

Based on RAI's documentation of the AI system's purpose, task, data, model, and relevant context, RAI conducted an issues identification exercise to surface specific risks and harms related to the use of AI systems for automated lending. As an example, the harms mapping surfaced a risk that a more opaque lending system may make it harder for a customer to identify errors related to product eligibility, amount, or pricing.

RAI validated identified issues with subject matter experts, including the Co-Chairs of RAI's Lending and Collections Working Group, past practitioners in the lending field, civil society groups, technical experts, lawyers, and AI engineers. Additionally, RAI received community input at a broader level from the members of the RAI Lending and Collections Working Group.

RAI's Certification Program for automated lending is based on a calibration of RAI Implementation Framework, which is described in the RAI Certification Program > Delivery section of this white paper. Based on the input from subject matter experts, RAI calibrated its Certification Assessment - including scoring and evidence requirements - for the automated lending use case.

In the meantime, based on its work with RAI, the FI has created a new governance structure for responsible AI, expanded its Model Risk Management (MRM) function to include responsible AI considerations, and adopted common processes and tools to support high-quality AI products that can scale.

RAI's Certification Program for automated lending is now formally under review by the Standards Council of Canada (SCC), prior to a harmonized review by SCC, the American National Standards Institute (ANSI), and United Kingdom Accreditation Service (UKAS). It is also being tested with additional FIs. When approved, it will be available for delivery to FIs via third-parties.

By May 2022, RAI had developed documentation for its Certification Program for automated lending, including a Certification Assessment, a Certification Guidebook, and a Scheme Guide. These materials are now being formally reviewed by RAI's Working Council and validated by the RAI community.

Context

Demand for a Global AI Certification Program

Given the wide variety of AI use cases and regulatory approaches, there is increasing demand from many quarters for a global AI certification program. Researchers have articulated the importance of certification programs to support good AI governance and to provide clear, actionable guidelines and instructions (Cihon et al., 2021; Dafoe, 2018; Leung, 2019; Marchant, 2019). The below table describes how RAI's Certification Program is designed to address the interests and concerns of stakeholders in the AI ecosystem.

Demand segments	Key stakeholders	Main interests/concerns
Suppliers	<ul style="list-style-type: none"> > Individual developers > Service providers/consulting firms > Suppliers of technology infrastructure 	<ul style="list-style-type: none"> > Knowing how to design and develop AI in a responsible way > Maximizing appropriate use and adoption of AI in a systematic and scalable way > Minimizing legal and business risk > Driving innovation and competitiveness > Differentiating themselves by having good processes in place > Increasing profitability and growth > Reducing operational costs
Buyers	<ul style="list-style-type: none"> > Procurement officers > Finance and legal teams > Senior management > Ethics boards and legal teams 	<ul style="list-style-type: none"> > Getting better procurement tools > Achieving business goals > Ensuring proper documentation, due diligence, and ethics

Demand segments	Key stakeholders	Main interests/concerns
Users	<ul style="list-style-type: none"> > Government decision makers > Individual consumers > Companies of all sizes 	<ul style="list-style-type: none"> > Reaping the benefits of AI (including by improving quality of life, changing behaviors, and taking better decisions) > Understanding what AI trustworthiness characteristics have been recognized internationally and how to evidence and measure them
End Users and Data Subjects	<ul style="list-style-type: none"> > Consumers and potential consumers > Employees and potential employees > People whose data/AI system uses 	<ul style="list-style-type: none"> > Ensuring fair and trustworthy functioning of AI systems > Ensuring privacy and security of data > Understanding what is being done to protect their interests and data
Educators and Researchers	<ul style="list-style-type: none"> > Academia > Educators > Research institutes 	<ul style="list-style-type: none"> > Educating the citizens and leaders of tomorrow > Disseminating tools, insights and knowledge
Lawmakers and public service	<ul style="list-style-type: none"> > National policy makers/regulators > Public sector 	<ul style="list-style-type: none"> > Minimizing harm to society > Increasing benefits of technology for humanity
Shapers	<ul style="list-style-type: none"> > UN > OECD > GPAI > G20 > Global AI Action Alliance (WEF) > Standards organizations > Industry associations > GAIA projects and partners* 	<ul style="list-style-type: none"> > Improving the state of the world by solving shared global challenges > Facilitating international and multi-stakeholder collaboration > Defining best practices for one or more industries > *Advancing the RAI agenda
Investors	<ul style="list-style-type: none"> > VCs > Trust funds > Pension funds > Philanthropies 	<ul style="list-style-type: none"> > Investing in quality AI systems that are fit for purpose > Answering demands for ethical investing > Maintaining profitability > Ensuring sustainability

Laws, Regulations, and Enforcement

October 2022

Global regulatory efforts to promote responsible AI adoption are gathering steam. The EU's proposed Artificial Intelligence Act, expected to be enacted in 2023, is the most ambitious such effort (European Commission, 2021). It employs a risk-based approach, heavily regulating systems that threaten fundamental human rights or safety (e.g. automated hiring, recidivism prediction). It also explicitly bans a further group of systems, including any that uses subliminal manipulation and any that engages in real-time biometric surveillance, except in specific cases.

In the US, the Federal Trade Commission (FTC) has announced plans to scrutinize—pursuant to the FTC Act, the Fair Credit Reporting Act, the Equal Opportunity Act—organizations that lack transparency, sufficient testing procedures, or quality datasets (Jillson, 2021; Smith, 2020; U.S. Equal Employment Opportunity Commission, 2021). The Department of Commerce's National Institute of Standards and Technology (NIST) put forth a draft AI Risk Management Framework (AI RMF) that seeks to showcase “what good looks like” for organizations deploying AI responsibly (NIST, 2022). RAI has engaged closely with NIST throughout the AI RMF development process and submitted



a comprehensive comment on the draft version (Responsible AI Institute, 2022). More recently, the US National AI Initiative office established the National AI Advisory Committee (NAIAC), chaired by Miriam Vogel, a member of RAII's Governing Board (National AI Initiative, 2022). The NAIAC is tasked with advising President Joe Biden and the National AI Initiative Office on topics related to the National AI Initiative.

More recently, the US National AI Initiative Office established the National AI Advisory Committee (NAIAC), chaired by Miriam Vogel, a member of RAII's governing board (National AI Initiative, 2022)

While more comprehensive proposed laws, like the 2022 Algorithmic Accountability Act, make their way through the American political process, federal agencies are undertaking ambitious AI-related initiatives. The Food and Drug Administration has announced plans for handling medical AI systems (FDA, 2021). State and municipal legislators have also taken steps to mitigate the risks of highly dangerous AI systems (Parker, 2021). For example, New York City has enacted a bill regulating the use of AI in employment contexts (New York City Council, 2021).

Though the EU is leading on AI-specific regulation, policymakers in the US and elsewhere - like UK, Canada, Japan, and Australia - are taking action to turn globally accepted AI principles into laws, regulations, and guidelines (Centre for Data Ethics and Innovation [CDEI], 2021; Department of Industry, Science, Energy and Resources, 2021; Government of Canada, 2021; Government of Ontario, 2021; Ministry of Economy, Trade & Industry, 2022). RAII's team tracks global AI laws and regulations carefully. RAII's Certification Program aligns with relevant laws and guidance and with the approaches proposed in comprehensive regulations, like the EU's AI Act. It is also informed by analyses of board responsibility for AI issues, relevant case law, and recent agency enforcement actions (Eccles & Vogel, 2022).

Globally Adopted AI Principles and Frameworks

The RAI Certification is grounded in Organisation for Economic Co-operation and Development (OECD) AI principles, which incorporate human rights objectives, good technology practices, and an emphasis on accountability and oversight (OECD, 2022). Additionally, the RAI Certification is informed by standards, guidelines, and other key principle and policy efforts, including, but not limited to, the following:

Document	Region	Relationship
UNESCO Recommendation on the Ethics of Artificial Intelligence (UNESCO, 2021)	International	RAI Implementation Framework is informed by UNESCO principles and framework to a significant degree, in areas such as data, governance, environment, gender, labor, and health.
IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (IEEE, 2021)	International	RAI Implementation Framework is informed by, covers, and provides further detail on the requirements of this under-development standard.
ISO proposed Artificial Intelligence Management Systems (ISO, 2021)	International	RAI Implementation Framework is informed by IEEE's Ethically Aligned Design principles to a significant degree, particularly in areas like explainability, transparency, and notification.
Global Partnership on AI (GPAI) Framework (GPAI, 2020)	International	RAI Implementation Framework incorporates concepts related to data governance, data rights, and data access from GPAI's framework paper on data governance.
World Economic Forum Procurement in a Box (WEF, 2020)	International	Ashley Casovan contributed to AI Procurement in a Box initiative, from which RAI Implementation Framework incorporates elements related to risk, governance, and procurement.

Document	Region	Relationship
Montreal Declaration for a responsible development of artificial intelligence (Université de Montréal, 2017)	International	RAI Implementation Framework incorporates principles and guidance from Montreal Declaration, in areas like transparency, fairness, notification, and safety.
NIST AI Risk Management Framework (NIST, 2022)	US	RAI Implementation Framework incorporates AI RMF requirements as they are developed. RAI team monitors and engages with NIST team developing AI RMF.
FTC guidance on AI (FTC, 2021)	US	RAI Implementation Framework is shaped to interoperate with the broad requirements outlined in FTC guidance.
OCC guidance on model risk management (OCC, 2021)	US	RAI Implementation Framework incorporates model risk management elements from OCC guidance.
FDA AI/ML-based Software as a Medical Device Action Plan (FDA, 2021)	US	RAI Implementation Framework is informed by FDA's ongoing translation of principles like transparency into specific requirements.
Canada's Directive on Automated Decision-Making Systems (Government of Canada, 2021)	Canada	Ashley Casovan led development of Canada's Directive, from which RAI Implementation Framework incorporates elements related to risk, governance, and procurement.
Office of the Superintendent of Financial Institutions Canada (OSFI) guidance (OSFI, 2020)	Canada	RAI Implementation Framework incorporates and expands upon OSFI's principles and requirements for soundness, explainability, and accountability.
EU Ethics guidelines for trustworthy AI (European Commission, 2019)	Europe	RAI Implementation Framework incorporates elements from EU Ethics guidelines, in areas like transparency, recourse, and bias.
Council of Europe's Report on AI systems (Council of Europe, 2020)	Europe	RAI Implementation Framework incorporates elements from Council of Europe Report, in areas like recourse, training, and transparency.

Document	Region	Relationship
The British Standards Institution (BSI) AI standards (BSI, 2022)	UK	RAI Implementation Framework is informed by BSI's understanding of how effective governance standards can promote privacy and protect consumers.
ICO Guidance on AI and data protection (ICO, 2020)	UK	RAI Implementation Framework is shaped to interoperate with the broad requirements outlined in ICO guidance.
UK Centre for Data Ethics and Innovation Roadmap to an effective AI assurance ecosystem (CDEI, 2021)	UK	RAI Implementation Framework incorporates the CDEI Roadmap's understanding of risk assurance roles, functions, and requirements.



Practitioner and Research Insights

October 2022

RAI's calibration of the global RAI Implementation Framework to specific use cases is informed by insights from practitioners and researchers. These include:

- > Practitioners currently developing an AI system for the use case
- > Practitioners who have previously developed AI systems for the use case
- > Researchers at the intersection of AI and the use case
- > Experts qualified to address complex questions related to harms, mitigation, and implementation
- > Industry organizations for implicated industries and functions
- > Advocacy organizations for people potentially impacted
- > Interested community members
- > Data scientists and AI/ML engineers
- > Legal and policy researchers
- > Responsible AI Working Groups on Automated Lending and Collections, Automated Skin Disease Detection, and/or Automated Human Resources
- > Responsible AI Working Council (arm's length internal approval body)

RAI Certification Program

October 2022

Scope

Recognizing that the term AI can have a variety of meanings, referring to many different types of technologies and tools, it is difficult to have a single certification program for all AI systems. While the same set of requirements should always be reviewed, it is important to consider responsible AI issues in the context of an AI system's use case, industry, and region. RAI's initial focus is on the following use cases:

- > Automated lending (Finance)
- > Automated collections (Finance)
- > Procurement (All Industries)
- > Human resources (All Industries)
- > Access to health care (Health Care)
- > Skin imaging (Health Care)

RAI's Certification Program for automated lending is now formally under review by the Standards Council of Canada (SCC), prior to a harmonized review by SCC, the American National Standards Institute (ANSI), and United Kingdom Accreditation Service (UKAS).

While the intent is for the certification to be globally adopted and expand to several more use cases, it has been important to focus on a few key areas to increase adoption.

Delivery

The Responsible AI Certification Program will be focused on the system level and delivered by a third-party organization. Based on the Responsible AI Implementation Framework, the certification assessment will:

1. Assess the data, model, and contextual deployment of an AI system, as these impact the efficacy, fairness, or usefulness of the system.
2. Use a set of 89 questions, response indicators, and evidence requirements to evaluate responsible AI maturity at the system level.
3. Consider the interplay of an AI system's domain, region, and system type.
4. Classify responsible AI considerations along the six responsible AI implementation dimensions (described below) and 20 responsible AI sub-dimensions (also called implementation requirements).
5. Provide detailed maturity scores for the AI system at the dimension and sub-dimension levels, which will determine the certification level that can be attributed to an AI system. See the graphic on the next page for a glimpse of the Certification Score Report format.





Assessment questions are generally scored on the following rubric:

Table 1. Scoring Rubric

Score	Description
0	Needs Improvement
1	Satisfactory
3	Good
5	Excellent

If an AI system earns 50%+ of available score in each dimension, each dimension score is totaled to get the total assessment score. This total assessment score is then represented as a percentage (total assessment score earned/total assessment score available). The assessment score percentage is used to determine the AI system’s certification level. The below table includes assessment score percentages and their corresponding certification levels:

Table 2. Certification Percentage, Level, and Logo

Total Score	Level Obtained	Corresponding Mark
0-49.9%	Not Certified	N/A
50-59.9%	Certified	
60-69.9%	Silver	 SILVER
70-79.9%	Gold	 GOLD
80+%	Platinum	 PLATINUM

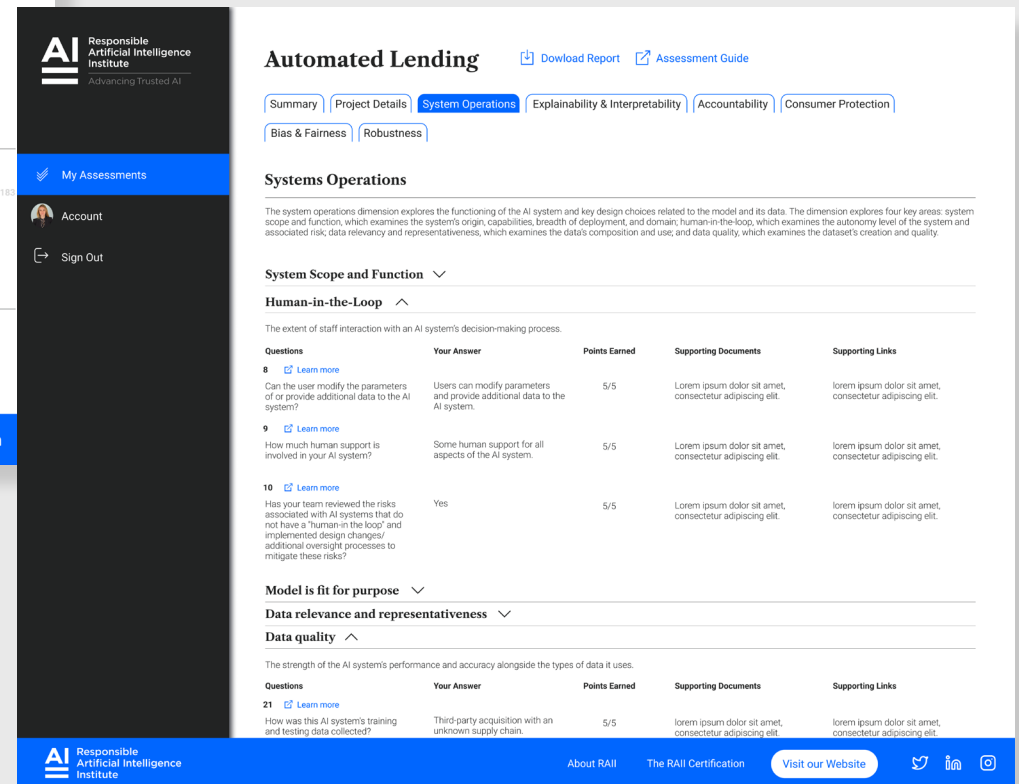
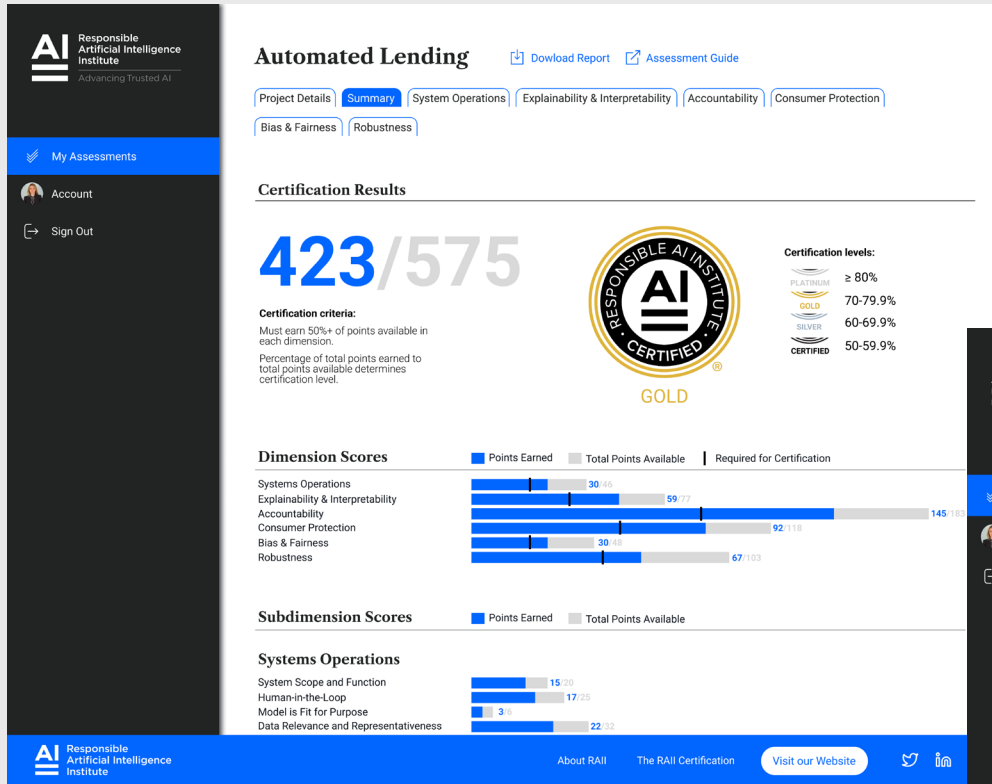


Figure 1. Sample of the Certification Score Report

Dimensions and Subdimensions

October 2022



Dimensions	Subdimensions
Data and System Operations	<ul style="list-style-type: none"> > System Scope and Function > Human-in-the-Loop > Model is Fit for Purpose > Data Relevance and Representativeness > Data Quality
Explainability & Interpretability	<ul style="list-style-type: none"> > Communication About the Outcome > Notification > Recourse > Understanding the AI System’s Decisions or Functions
Accountability	<ul style="list-style-type: none"> > Organizational Governance > Team Governance
Consumer Protection	<ul style="list-style-type: none"> > Transparency to the User and Data Subject > Harms to Individuals > Protections
Bias & Fairness	<ul style="list-style-type: none"> > Bias Impacts > Bias Training > Bias Testing
Robustness	<ul style="list-style-type: none"> > Data Drift > System Acceptance Test is Performed > Contingency Planning

Data and System Operations

The data and system operations dimension explores the functioning of the AI system and key design choices related to the model and its data.

The dimension explores four key areas: system scope and function, which examines the system's origin, capabilities, breadth of deployment, and domain; human-in-the-loop, which examines the autonomy level of the system and associated risk; data relevancy and representativeness, which examines the data's composition and use; and data quality, which examines the dataset's creation and quality (Berendt & Preibusch, 2014; Demartini et al., 2017; Jotter & Bosco, 2020).

Explainability and Interpretability

The explainability and interpretability dimension ensures that the AI system's workings and uses can be explained and documented in terms that humans - including users, data subjects, and others - can understand. This involves inspecting the complexity of the system - like its capabilities, how it was trained - plus any steps taken by the team to bolster the system's explainability (like prioritizing simple models during the design process, implementing integration tests to understand how individual components interact with each other). It also involves analyzing how the system presents information to its users and data subjects: how it communicates the outcome and the reasoning behind that outcome, whether it provides notification that an AI system was involved in the generation of that outcome, and whether it offers and communicates opportunities for redress.

Accountability

The accountability dimension examines whether the organization has set up clear oversight processes for the development and implementation of the

AI system. These oversight processes should ensure that the organization is held accountable for designing a system that is explainable, fair, and not manipulative, as well as for clearly communicating the system's functions and limitations to its users. The accountability dimension also verifies that the AI system development team has documented design choices, reviewed system failures, and conducted an appropriate scenario planning exercise.

Consumer Protection

The consumer protection dimension evaluates the risk the AI system poses to individuals and the steps the organization and development team have taken to mitigate these risks. The assessment studies transparency - whether data policies, system risks, testing results, and appropriate uses are communicated to users and data subjects. It also estimates the maximum potential harm of the AI system and checks whether the team has completed appropriate mitigation exercises such as harms mapping and root cause analysis. The assessment is also concerned with privacy, cataloging what sensitive data (like personal data, demographic information, or business data) is used during training and deployment, and what strategies the team has employed to protect that data.

Bias and Fairness

The bias dimension assesses whether the AI system was designed in a manner that promotes fairness and avoids bias. The extent to which the organization and development team have engaged with bias and fairness issues, such as by conducting research, situating the system in its historical and cultural context, hiring team members with relevant expertise, and providing opportunities for workers displaced by the system, is considered. The assessment also reviews any bias training that the organization has provided to the AI system's

users. Finally, the team's testing procedures are analyzed: tests that employ appropriate fairness definitions and that consider multiple types of potential bias should be performed on an ongoing basis (WEF, 2018).

Robustness

The robustness dimension investigates if the AI system is safe and effective. Its questions ascertain whether the system is adequately protected against data drift, as well as whether it is robust enough to handle edge cases and extreme scenarios. This dimension also checks what testing, like accuracy tests or unit tests, are completed and at what frequency.

RAI Community

October 2022

Ecosystem

Our Team

Leadership Team

- > Ashley Casovan, Executive Director
- > Var Shankar, Director, Policy, Delivery and Member Success
- > Benjamin Faveri, Research and Policy Analyst
- > Amanda Lawson, Research and Policy Analyst
- > Ifejesu Ogunleye, Research and Policy Analyst
- > Alyssa Lefavre, Director, Partnerships and Market Development

Governing Board

- > Manoj Saxena, Executive Chairman
- > Miriam Vogel, President and CEO, EqualAI
- > Michael Stewart, Founder, Chairman, and CEO, Lucid.AI
- > Matt Sanchez, Founder and CTO, CognitiveScale
- > Joydeep Ghosh, Schlumberger Centennial Chair Professor of Electrical and Computer Engineering at The University of Texas at Austin

Working Council Chair

- > Craig Shank, Principal at CES.world and Former Vice-President of Standards at Microsoft

Lending & Collections Working Group Co-Chairs

- > Aurelie Jacquet, Ethical AI Consulting
- > Suraj Madhani, American Express

Human Resources Working Group Co-Chairs

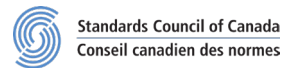
- > Matissa Hollister, McGill University

- > Barbara Cosgrove, Workday

RAI's Corporate Members



NGOs & Standard Bodies



Academia & Government



Industry Collaborators



Acknowledgements

October 2022

RAI thanks the following individuals for contributing to RAI's approach as described in this white paper:

- > Kasia Chmielinski, Data Nutrition Project
- > Tim Vincent, IBM
- > Julia Stoyanovich, New York University
- > Brandie M. Nonnecke, CITRIS Policy Lab, University of California, Berkeley
- > Umang Bhatt, PhD Candidate, Machine Learning Group, University of Cambridge
- > Nandita Sampath, Consumer Reports
- > Krista Kinnard, U.S. Department of Labor
- > Alka Patel, Former Head of AI Ethics Policy for the Joint AI Center at the U.S. Department of Defense
- > Veronica Rotemberg, Memorial Sloan Kettering Cancer Center
- > Jessica Newman, AI Security Initiative, University of California, Berkeley
- > Jayant Narayan, World Economic Forum Global AI Action Alliance

RAI also thanks Hannah Brooks, Stephanie Cairns, Tina Lassiter, and Anders Liman for their work on this white paper in its current and/or prior versions.

References

Andrew Smith. 2020. Using Artificial Intelligence and Algorithms. <https://www.ftc.gov/news-events/blogs/business-blog/2020/04/using-artificial-intelligence-algorithms>

Bettina Berendt & Sören Preibusch. 2014. Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence. *Artificial Intelligence and Law*. 22(2). 175-209. <https://doi.org/10.1007/s10506-013-9152-0>

BSI. 2022. Artificial Intelligence. <https://standardsdevelopment.bsigroup.com/committees/50281655>

Centre for Data Ethics and Innovation. 2021. The roadmap to an effective AI assurance ecosystem. <https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem>

Council of Europe. 2020. Towards Regulation of AI Systems: Global perspectives on the development of a legal framework on Artificial Intelligence (AI) systems based on the Council of Europe's standards on human rights, democracy and the rule of law. <https://rm.coe.int/prems-107320-gbr-2018-compli-cahai-couv-texte-a4-bat-web/1680a0c17a>

Dafoe, A. 2018. AI Governance: A Research Agenda. Future of Humanity Institute, University of Oxford, Center for the Governance of AI. <https://www.fhi.ox.ac.uk/govai/>

Department of Industry, Science, Energy and Resources. 2021. Australia's Artificial Intelligence Ethics Framework. <https://www.industry.gov.au/data-and-publications/australias-artificial-intelligence-ethics-framework>

Elisa Jillson. 2021. Aiming for truth, fairness, and equity in your company's use of AI. <https://www.ftc.gov/news-events/blogs/business-blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai>

European Commission. 2019. Ethics guidelines for trustworthy AI. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

European Commission. 2021. 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. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206>

Fabrice Jotter and & Clara Bosco. 2020. Keeping the “Human in the Loop” in the Age of Artificial Intelligence. *Science and Engineering Ethics*. 26(5). 2455-2460. <https://doi.org/10.1007/s11948-020-00241-1>

FDA. 2021. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. <https://www.fda.gov/media/145022/download>

FTC. 2021. Aiming for truth, fairness, and equity in your company’s use of AI. <https://www.ftc.gov/business-guidance/blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai>

Gianluca Demartini, Djelle Eddine Difallah, Ujwal Gadiraju, & Michele Catasta. 2017. An Introduction to Hybrid Human-Machine Information Systems. *Foundations and Trends in Web Science*. 7(1). 1-87. <https://doi.org/10.1561/18000000025>

Government of Canada. 2021. Directive on Automated Decision-Making. <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592>

Government of Canada. 2021. Responsible use of artificial intelligence (AI). <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai.html>

Government of Ontario. 2021. Consultation: Ontario’s Trustworthy Artificial Intelligence (AI) Framework. <https://www.ontario.ca/page/ontarios-trustworthy-artificial-intelligence-ai-framework-consultations>

GPAI. 2020. A Framework Paper for GPAI’s work on Data Governance. <https://gpai.ai/projects/data-governance/gpai-data-governance-work-framework-paper.pdf>

Grand View Research. 2022. Artificial Intelligence In Healthcare Market Size, Share, And Trends Analysis Report By Component (Software Solutions, Hardware, Services), By Application (Virtual Assistants, Connected Machines), By Region, And Segment Forecasts, 2022 - 2030. <https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-healthcare-market>

ICO. 2020. Guidance on AI and data protection. <https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-artificial-intelligence-and-data-protection/>

ICO. 2021. Certification. <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/certification/>

IEEE. 2021. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. <http://standards.ieee.org/industry-connections/ec/autonomous-systems/>

ISO. 2021. ISO/IEC CD 42001. <https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/08/12/81230.html>

Jade Leung. 2019. Who will govern artificial intelligence? Ph.D. Dissertation. University of Oxford. <https://ora.ox.ac.uk/objects/uuid:ea3c7cb8-2464-45f1-a47c-c7b568f27665>

Jake Parker. 2021. Most State Legislatures Have Rejected Bans and Severe Restrictions on Facial Recognition. <https://www.securityindustry.org/2021/07/09/most-state-legislatures-have-rejected-bans-and-severe-restrictions-on-facial-recognition>

Marchant, G. 2019. "Soft Law" Governance of Artificial Intelligence. Arizona State University, Center for Law, Science & Innovation. <https://aipulse.org/soft-law-governance-of-artificial-intelligence/>

Ministry of Economy, Trade & Industry. 2022. Governance Guidelines for Implementation of AI Principles Ver. 1.1. https://www.meti.go.jp/english/press/2022/0128_003.html

National AI Initiative. 2022. The National AI Advisory Committee (NAIAC). https://www.ai.gov/naiac/#ABOUT_THE_ADVISORY_COMMITTEE

New York City Council. 2021. Automated Employment Decision Tools. <https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9>

Nicol Turner Lee & Samantha Lai. 2021. Why New York City is cracking down on AI in hiring. Brookings Institution. <https://www.brookings.edu/blog/techtank/2021/12/20/why-new-york-city-is-cracking-down-on-ai-in-hiring/>

NIST. 2022. AI Risk Management Framework. <https://www.nist.gov/system/files/documents/2022/03/17/AI-RMF-1stdraft.pdf>

OCC. 2021. Model Risk Management. <https://www.occ.gov/publications-and-resources/publications/comptrollers-handbook/files/model-risk-management/index-model-risk-management.html>

OECD. AI. 2022. The OECD Artificial Intelligence (AI) Principles. <https://oecd.ai/en/ai-principles>

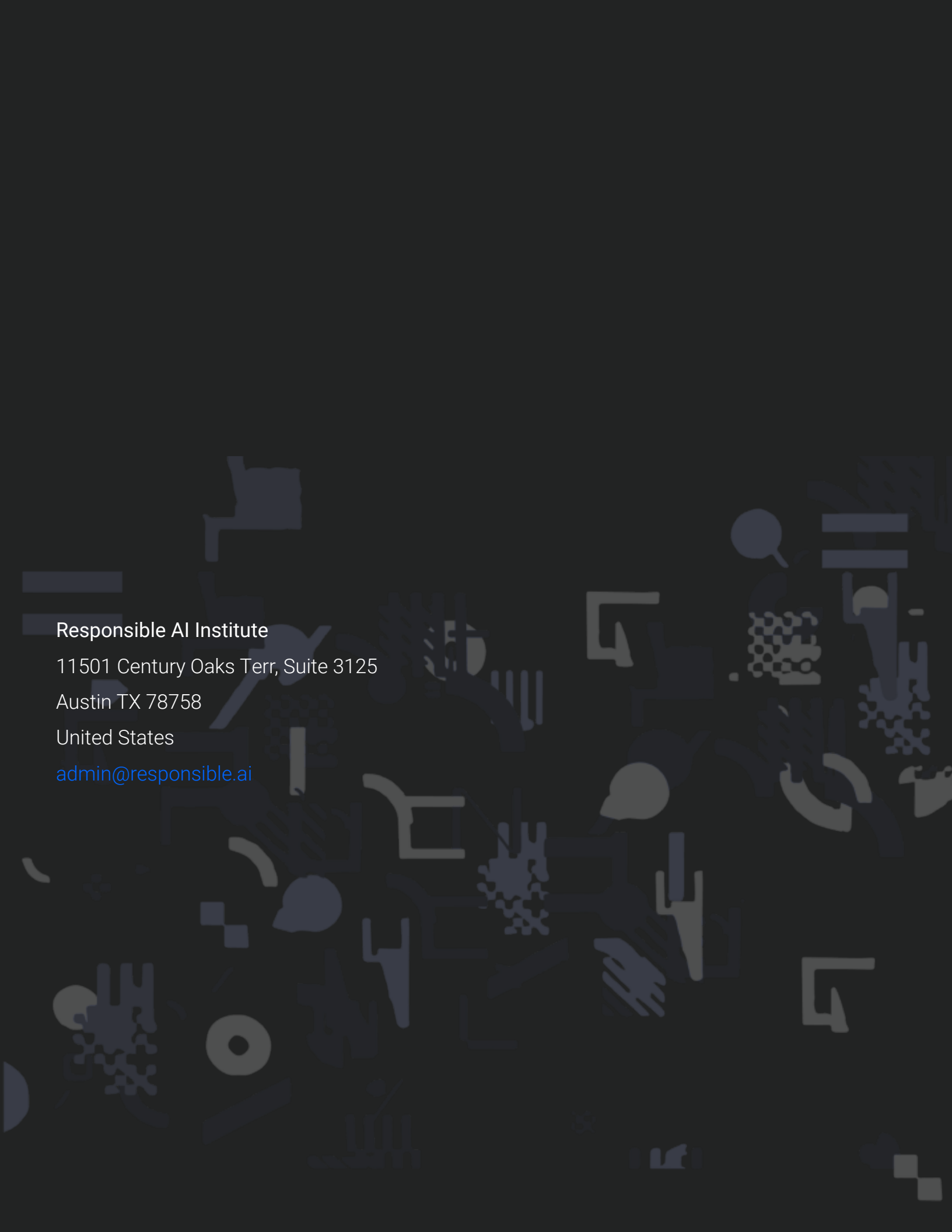
- OSFI. 2020. Developing Financial Sector Resilience in a Digital World. <https://www.osfi-bsif.gc.ca/Eng/Docs/tchrsk.pdf>
- Peter Cihon. 2019. Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development. Future of Humanity Institute, University of Oxford, Center for the Governance of AI. 1-41.
- Peter Cihon, Matthijs M. Maas, and Luke Kemp. 2020. Should Artificial Intelligence Governance be Centralised? Design Lessons from History. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). Association for Computing Machinery, New York, NY, USA, 228-234. <https://doi.org/10.1145/3375627.3375857>
- Peter Cihon, Moritz J. Kleinaltenkamp, Jonas Schuett, & Seth D. Baum. 2021. AI Certification: Advancing Ethical Practice by Reducing Information Asymmetries. IEEE Transactions on Technology and Society 2, 4 (Dec. 2021), 200-209. <https://doi.org/10.1109/TTS.2021.3077595> arXiv: 2105.10356.
- Responsible AI Institute. 2022. Comment from the Responsible AI Institute on the Initial Draft of the NIST AI Risk Management Framework. <https://www.nist.gov/system/files/documents/2022/05/19/Responsible%20AI%20-%20Comments.pdf>
- Robert Eccles & Miriam Vogel. 2022. Board Responsibility for Artificial Intelligence Oversight. <https://corpgov.law.harvard.edu/2022/01/05/board-responsibility-for-artificial-intelligence-oversight/>
- UNESCO. 2021. Recommendation on the ethics of artificial intelligence. <https://unesdoc.unesco.org/ark:/48223/pf000038045>
- Université de Montréal. 2017. Montreal Declaration for a Responsible Development of Artificial Intelligence. <https://www.montrealdeclaration-responsibleai.com/process>
- U.S. Equal Employment Opportunity Commission. 2021. EEOC Launches Initiative on Artificial Intelligence and Algorithmic Fairness | U.S. Equal Employment Opportunity Commission. <https://www.eeoc.gov/newsroom/eeoc-launches-initiative-artificial-intelligence-and-algorithmic-fairness>
- Valeria Marcia & Kevin C. Desouza. 2021. The EU path towards regulation on artificial intelligence. Brookings Institution. <https://www.brookings.edu/blog/techtank/2021/04/26/the-eu-path-towards-regulation-on-artificial-intelligence/>

World Economic Forum. 2018. How to Prevent Discriminatory Outcomes in Machine Learning. https://www3.weforum.org/docs/WEF_40065_White_Paper_How_to_Prevent_Discriminatory_Outcomes_in_Machine_Learning.pdf

World Economic Forum. 2020. AI Procurement in a Box. <https://www.weforum.org/reports/ai-procurement-in-a-box/>

World Economic Forum. 2021. Human-Centred AI for HR: State of Play and the Path Ahead. https://www3.weforum.org/docs/WEF_Human_Centred_AI_for_HR_2021.pdf

World Economic Forum. 2021. World Economic Forum Launches New Global Initiative to Advance the Promise of Responsible Artificial Intelligence. <https://www.weforum.org/press/2021/01/world-economic-forum-launches-new-global-initiative-to-advance-the-promise-of-responsible-artificial-intelligence/>



Responsible AI Institute

11501 Century Oaks Terr, Suite 3125

Austin TX 78758

United States

admin@responsible.ai