A Certification for Responsible AI

February 1, 2022

Responsible
Artificial Intelligence
Institute

Table of contents

Background

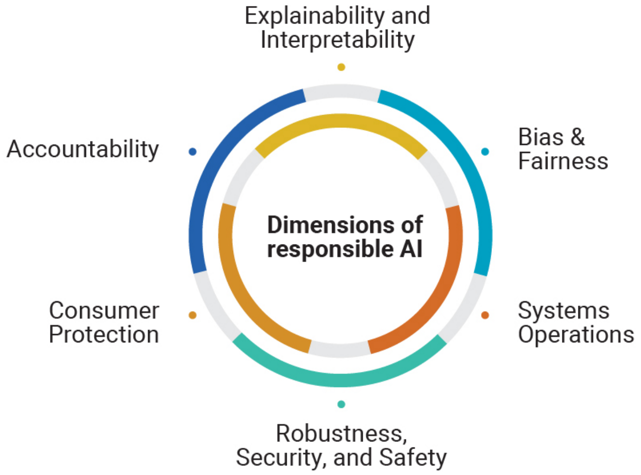
About RAII

The Responsible AI Institute (RAII) is developing one of the world's first responsible AI certification programs grounded in human rights and aligned with globally adopted AI principles, research, and emerging best practices and regulations. RAII is an independent and community-driven non-profit organization building tangible governance tools for trustworthy, safe, and fair AI.

RAII is a member of the World Economic Forum (WEF) Global AI Action Alliance (GAIA), which comprises over 100 government entities, civil society organizations, private companies, and academic institutions dedicated to the responsible use of AI. GAIA members work together to identify and implement tools and best practices that promote ethical AI.

The RAII Certification Program is a maturity assessment that evaluates AI systems. It tests both the risk of an AI system and the corresponding mitigation measures. Recognizing that not all AI systems are the same, this program is tailored to specific industries and use cases starting with health, finance, HR, and procurement.

RAII Implementation Framework Overview


Explainability and Interpretability
Accountability
Bias & Fairness
Dimensions of responsible AI
Consumer Protection
Systems Operations
Robustness, Security, and Safety

Sub-dimensions of responsible AI

Accountability

- ✓ Clear oversight process for implementation of AI
- ✓ Independent review process and ongoing monitoring

Bias & Fairness

- ✓ Human rights/ethics acceptance
- ✓ Bias training and education
- ✓ Test for unwanted bias

Robustness, Security, and Safety

- ✓ Data drift
- ✓ System Acceptance Test
- ✓ Contingency planning

Systems Operations

- ✓ System scope and function
- ✓ Human-in-the-loop
- ✓ Model is fit for purpose
- ✓ Representative and relevant data
- ✓ Data quality

Consumer Protection

- ✓ Transparency to the user and data subject
- ✓ Harm to individuals/incident reporting
- ✓ System protects individual's or group's privacy

Explainability and Interpretability

- ✓ Communication about the outcome
- ✓ Notification
- ✓ Recourse
- ✓ Clear understanding of how the system arrives at a decision or function

The RAII Certification

For several years, governments, companies, and civil society organizations have grappled with how to govern AI systems in a consistent manner. As the number of organizations putting forward principles has multiplied, an international consensus has emerged on what constitutes responsible AI. These efforts led to the adoption of principles by organizations around the world. These principles are now informing a burgeoning and expansive set of standards and regulations.

The challenge lies in implementation and adoption. The RAII Certification translates the globally adopted principles, standards, and current regulations into clear implementation requirements taking the guesswork out of what it means to be responsible.

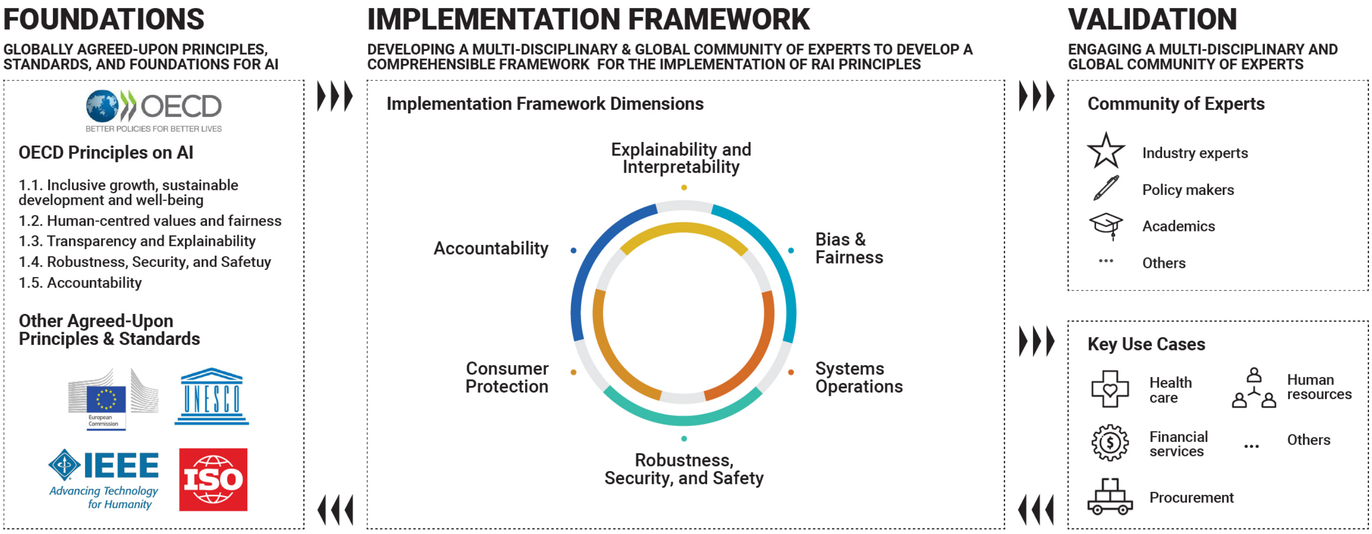
Informed by those researching, designing, building, deploying, using, and overseeing AI, we have aggregated all of this information to understand:

- What is responsible AI - common objectives, uses, and definitions
- Demand signals for oversight - who is concerned, who are the key players, and what are their main interests or concerns
- How certification can support responsible AI adoption

Developing a Common Definition of Responsible AI

RAII and its board of advisors have been at the forefront of AI development and policy and governance of AI systems. In addition to their lived experience and firsthand challenges, the team has spent the last three years integrating leading research, learnings from discussions with leading AI policy makers, and results of use case tests with those designing, developing, and using AI systems into a comprehensive framework, which includes a high-level set of implementation objectives and requirements.

Additional details about the certification categories and how they are assessed can be found in the section “RAII Implementation Framework Details” on page 9.



FOUNDATIONS

GLOBALLY AGREED-UPON PRINCIPLES, STANDARDS, AND FOUNDATIONS FOR AI

OECD Principles on AI

1.1. Inclusive growth, sustainable development and well-being
1.2. Human-centred values and fairness
1.3. Transparency and Explainability
1.4. Robustness, Security, and Safetuy
1.5. Accountability

Other Agreed-Upon Principles & Standards

OECD
BETTER POLICIES FOR BETTER LIVES
European Commission
UNESCO
IEEE
Advancing Technology for Humanity
ISO

IMPLEMENTATION FRAMEWORK

DEVELOPING A MULTI-DISCIPLINARY & GLOBAL COMMUNITY OF EXPERTS TO DEVELOP A COMPREHENSIBLE FRAMEWORK FOR THE IMPLEMENTATION OF RAI PRINCIPLES

Implementation Framework Dimensions

Explainability and Interpretability
Bias & Fairness
Systems Operations
Robustness, Security, and Safety
Consumer Protection
Accountability

VALIDATION

ENGAGING A MULTI-DISCIPLINARY AND GLOBAL COMMUNITY OF EXPERTS

Community of Experts

- Industry experts
- Policy makers
- Academics
- Others

Key Use Cases

- Health care
- Human resources
- Financial services
- Others
- Procurement

AI Certification Demand Signals

The development of the RAII Certification Framework and its application is in response to increasingly significant demand signals. While leading academic researchers have articulated the importance of certification programs to support good AI governance, there are important drivers being articulated from several key audiences involved in the AI adoption life-cycle. The table below outlines the key takeaways from RAII's research and engagement activities.

Demand segments	Key stakeholders	Main interests/concerns
Those who develop	• Individual developers • Ethics boards and legal teams • Service providers/consulting firms • Suppliers of technology infrastructure	• Knowing how to design and develop AI in a responsible way • Minimizing business and legal risks • Maximizing use and adoption of AI • Driving innovation and competitiveness • Increasing profitability and growth • Reducing costs of doing business
Those who buy	• Procurement officers • Finance and legal teams • Senior management	• Getting better procurement tools • Achieving business goals • Ensuring proper due diligence and ethics
Those who use	• Government decision makers • Individual consumers • Companies of all sizes	• Reaping the benefits of AI (including by improving quality of life, changing behaviors, and taking better decisions)
Those who educate and research	• Academia • Educators • Research institutes	• Educating the citizens and leaders of tomorrow • Disseminating tools, insights and knowledge
Those who control and regulate	• National policy makers and regulators • Standards organizations • Industry associations	• Minimizing harm to society • Increasing benefits of technology for humanity • Protecting the stakeholders of an industry
Those who shape	• UN • OECD • GPAI • G20 • WEF • GAIA projects and partners*	• Improving the state of the world by solving shared global challenges • Facilitating international and multi-stakeholder collaboration • *Advancing the RAI agenda
Those who invest	• VCs • Trust funds • Pension funds • Philanthropies	• Answering demands for ethical investing • Maintaining profitability • Ensuring sustainability

Scope of RAII Certification Program

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. RAII's initial focus in on the following use cases:

All industries	Finance	Health
• **AI Procurement** • HR systems	• **Automated lending** • Automated collections	• Access to care • **Skin imaging**

The use cases in bold above are in the process of being reviewed as part of a formal certification review process submitted to national accreditation bodies in the United States, Canada, and the United Kingdom.

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.

Globally Adopted AI Principles

The RAII Certification is grounded in OECD's AI principles as they are the most established AI principles, incorporating human rights objectives, good technology practices, and an emphasis on accountability and oversight.

Additionally, the RAII Certification is informed by standards, guidelines, and other key principle and policy efforts, including, but not limited to, the following:

- UNESCO Recommendation on the Ethics of Artificial Intelligence
- IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems
- ISO proposed Artificial Intelligence Management Systems
- NIST AI Risk Management Framework
- FTC guidance on AI
- OCC guidance on model risk management
- FDA AI/ML-based Software as a Medical Device Action Plan
- Canada's Directive on Automated Decision-Making Systems
- OSFI guidance
- EU Ethics guidelines for trustworthy AI
- Council of Europe's Report on AI systems
- BSI AI standards
- UK certification guidelines
- Global Partnership on AI Framework
- WEF Procurement in a Box

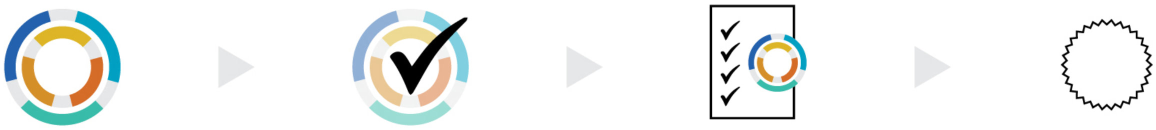
Existing and Upcoming Laws and Regulations

In classifying and assessing AI systems according to their risk levels, RAII's approach aligns with the EU's proposed AI Act, GDPR, and various digital and data regulations - both existing and upcoming - from other regions.

In the US, RAII closely follows AI-related guidance from the Federal Trade Commission, the White House, and the National Security Commission on Artificial Intelligence, as well as AI- and industry-specific AI regulations at the federal, state, and local levels.

Certification Approach

The RAII Certification is based on research, member engagements, workshops, and expert guidance and validated by a RAII council of ethics, technical, and legal experts. Getting to certification is a process comprised of the following four stages:



Build

A comprehensive community-validated framework

Accredit

Framework assessed by National accreditation bodies

Audit

Audits performed by trained and independent auditors

Certify

A responsible AI certification is awarded

Industry and Academic Research

As interest in AI has grown exponentially in the past few years, academic and industry research has become increasingly specialized. So, when developing assessment tools for any use case, RAII works with leading researchers and practitioners to integrate the latest academic and industry research relevant to each use case - both at the outset and on an ongoing basis. A full list of RAII's advisors and academic partners is available on our website.

Member Engagements

RAII's engagements with its member organizations inform its assessment tools. These member engagements allow RAII and member organizations to rapidly integrate the latest relevant research into business processes, while surfacing implementation issues that require further expert guidance or research.

Expert Guidance

Subject matter experts inform each RAII assessment in diverse ways. RAII Working Groups - each of which convenes experts around a use case - include functionally diverse members from industry, academia, and civil society, so that AI issues related to the use case can be examined from multiple angles. RAII also directly engages specific experts on relevant questions.

RAII Council

Each RAII assessment is validated by RAII's Council of Advisers, which includes ethics, technical, and legal experts.

Testing the Certification with Real Use Cases

A key goal of the RAII Certification is to help bridge the gaps between responsible AI practitioners, researchers, and regulators. It is informed by practitioner insights, industry and academic research, and existing and upcoming regulation. RAII's initial focus is on use cases that are currently available in the market and have the potential to cause significant harm if not used appropriately.

Health care

AI systems within the healthcare industry are currently valued at $6.7 billion. RAII worked with a leading American health insurer to develop policies, governance, and assessment tools to develop and scale its responsible AI program. Through this effort, RAII discovered a need for the organization to adopt AI-specific governance practices, build capacity for bias testing, update and improve data collection practices, and introduce role-relevant responsible AI training in order to protect their organization and their clients from unintended consequences when using AI systems. When these improvements are made, the health insurer will be able to demonstrate compliance with future AI regulations. RAII's initial health care use cases include automated pre-authorization for health insurance and the use of computer vision techniques to diagnose skin disease.

Finance

The use of AI systems in the financial services industry is likely to increase exponentially in the coming years. McKinsey estimates that AI systems could eventually deliver $1 trillion of annual market value. In preparation for a future certification and regulations, RAII worked with a Canadian bank to evaluate the responsibility of three AI systems related to lending decisions, product recommendation, and optical character recognition. In doing so, RAII found both common challenges and different potential harms related to each use case, further articulating the need for use-case specific certification programs.

RAII's Automated Lending and Collections Working Group discussed many key findings and best practices to mitigate potential harms resulting from these and related systems.

Procurement

Building from our work with WEF's Procurement in a Box initiative, RAII is of the conviction that organizations need to test and verify the AI tools they are procuring in the same way that they govern the development and use of AI systems they build, while acknowledging that there are responsibilities and potential liabilities on both parties. To understand how an AI Certification for procurement can be leveraged to protect both organizations buying and selling AI systems, RAII has been working with a government agency on a AI procurement pilot, exploring what evaluations are required at which stage in the procurement lifecycle, and examining what type of training is needed to assist current procurement specialists.

Benefits of Certification by Key Audience

Key Audiences	Value of RAII Certification for Audience Group
Senior Executives & Executive Review Boards	Confidence they and their teams have taken all necessary precautions to mitigate bias within their AI systems.
Compliance Officers	Consistent reporting and information sharing between first and second line of defence within a company.
Procurement Officers	Knowledge they are procuring trustworthy AI systems that won't cause the organization liability or risk.
Regulators	Compliance with established regulations.
Investors	Assurance they're investing in AI systems that will mitigate harm to people and the planet.
Consumers	Comfort knowing their rights, privacy, and civil liberties are protected.

RAII Implementation Framework Details

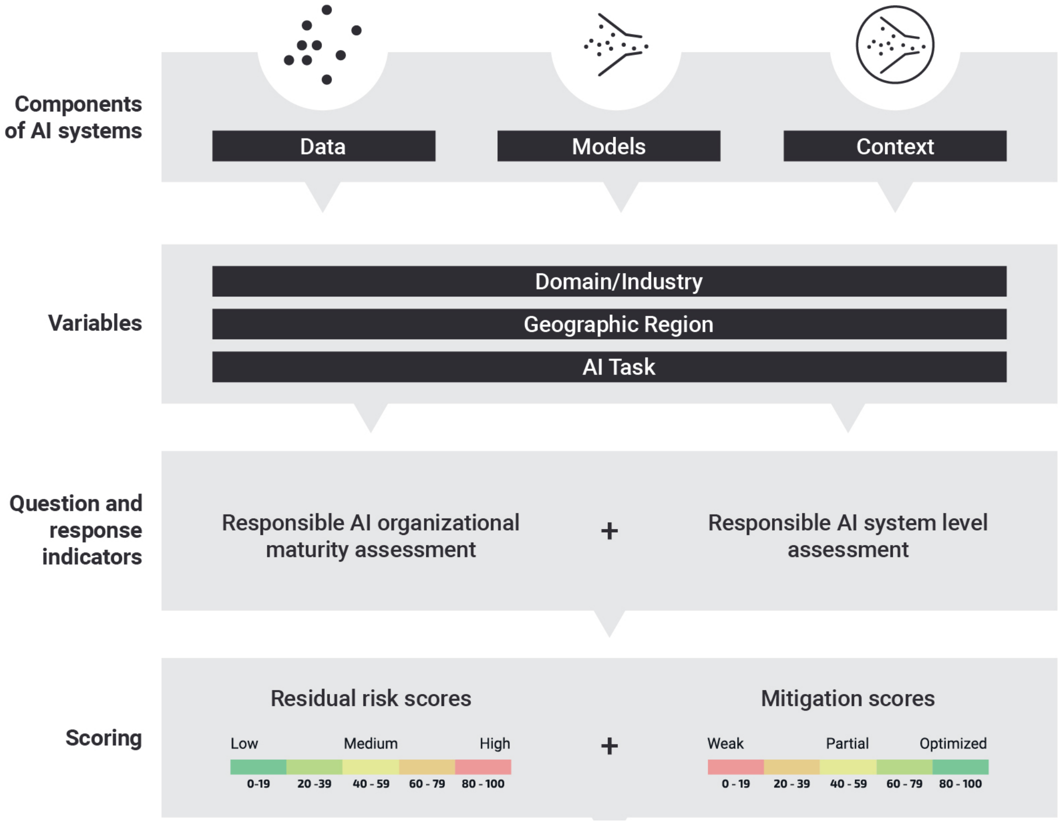
Informed by globally-agreed upon principles and by industry and academic research, RAII has identified foundational responsible AI principles, practices, and concepts with which to build a strong implementation framework for the RAII Certification and other RAII tools.

The RAII implementation framework:

1. Assesses the data, model, and contextual deployment of an AI system, as these are all factors impacting the efficacy, fairness, or usefulness of the system.
2. Considers the interplay of an AI system's domain, region, and system type.
3. Classifies responsible AI considerations along six responsible AI implementation concepts and 24 responsible AI implementation requirements.
4. Uses a set of 200+ questions and response indicators to evaluate responsible AI maturity at the organizational and system levels.
5. Provides detailed residual risk and mitigation maturity scores for the AI system, which together determine the certification level that can be attributed to an AI system.

This implementation framework also informs the RAII Certification. The heart of the certification is an assessment that consists of risk and mitigation questions related to each of the six responsible AI implementation concepts and the 24 responsible AI implementation requirements.

The RAII Certification for a given use case - delivered independently by a third-party - will be awarded to an AI system if the assessment results for each implementation concept are adequate.



RAII Assessment™ Architecture

Framework Dimensions

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.

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.

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, cataloguing 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.

Robustness, Security, and Safety

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.

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.

System Operations

The 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.

Review and Feedback

The RAII Certification Program is currently under review by several of our members and advisors. A full list will be provided in future versions of this document. If you are interested in learning more about the certification program, participating in the review process, or learning more about RAII please contact us at admin@responsible.ai.

About RAII

Members and Partners

Corporate Members

IBM Anthem AltaML Yum! JACKSON

ATB Financial ARMILLA CognitiveScale YOUR COGNITIVE CLOUD COMPANY

NGOs & Standard Bodies

WORLD ECONOMIC FORUM Standards Council of Canada Conseil canadien des normes IEEE Advancing Technology for Humanity The ROCKEFELLER FOUNDATION UNESCO OECD BETTER POLICIES FOR BETTER LIVE

ALGORA ISO DATA NUTRITION PROJECT ceimia SAXENA FOUNDATION OCEANIS OPEN COMMUNITY FOR ETHICS IN AUTONOMOUS AND INTELLIGENT SYSTEMS

Academia & Government

TEXAS The University of Texas at Austin Centre for Data Ethics and Innovation CITRIS AND THE BANATAO INSTITUTE UNIVERSITY OF TORONTO UNIVERSITY OF CAMBRIDGE UNIVERSITY OF ALBERTA

OXFORD BROOKES UNIVERSITY UNIVERSITY OF MICHIGAN r/ai center for responsible ai Queen's UNIVERSITY Université de Montréal MICHIGAN STATE UNIVERSITY

PRINCETON UNIVERSITY NYU Canadä GAO U.S. GOVERNMENT ACCOUNTABILITY OFFICE SCHWARTZ REISMAN INSTITUTE FOR TECHNOLOGY AND SOCIETY DEPARTMENT OF DEFENSE UNITED STATES OF AMERICA DEPARTMENT OF LABOR UNITED STATES OF AMERICA

Industry Collaborators

Microsoft Oproma HSBC HYPER GIANT SKINOPATHY EY

Deloitte. BEACON AMERICAN EXPRESS HEXAGON RISING TO EVERY CHALLENGE Mila

RAII Team

Leadership and Team

Ashley Casovan, Executive Director

Var Shankar, Senior Manager, Policy, Delivery and Member Success

Hannah Brooks, Research and Policy Analyst

Benjamin Faveri, Research and Policy Analyst

Stephanie Cairns, Research and Policy Analyst

Phil Dawson, Senior Policy Counsel

Alyssa Lefaivre, Director, Partnerships and Market Development

Kara Scully, Manager, Communications and Engagements

Raidhy Hererra, Marketing and Development Analyst

Governing Board

Manoj Saxena, Executive Chairman

Michael Stewart, Founder, Chairman, and CEO, Lucid.AI

Matt Sanchez, Founder and CTO, CognitiveScale

Miriam Vogel, President and CEO, EqualAI

Joydeep Ghosh, Schlumberger Centennial Chair Professor of Electrical and Computer Engineering at The University of Texas at Austin

Responsible
Artificial Intelligence
Institute