Responsible
Artificial Intelligence
Institute

Advancing Trusted AI

Presentation to Nuclear Regulatory Commission

Var Shankar
Responsible AI Institute
September 19, 2023

Parts

01 RAI Institute’s Efforts and Certification

02 AI Regulation and Nuclear Regulation

03 CWG White Paper on AI Certification

1/3 RAI Institute’s Efforts and Certification

Responsible AI Institute

The Responsible AI Institute (RAI Institute) is dedicated to enabling successful responsible AI efforts in organizations.

Responsible Artificial Intelligence Institute
Advancing Trusted AI

Independent not–for–profit

Subject matter expertise

Member–driven

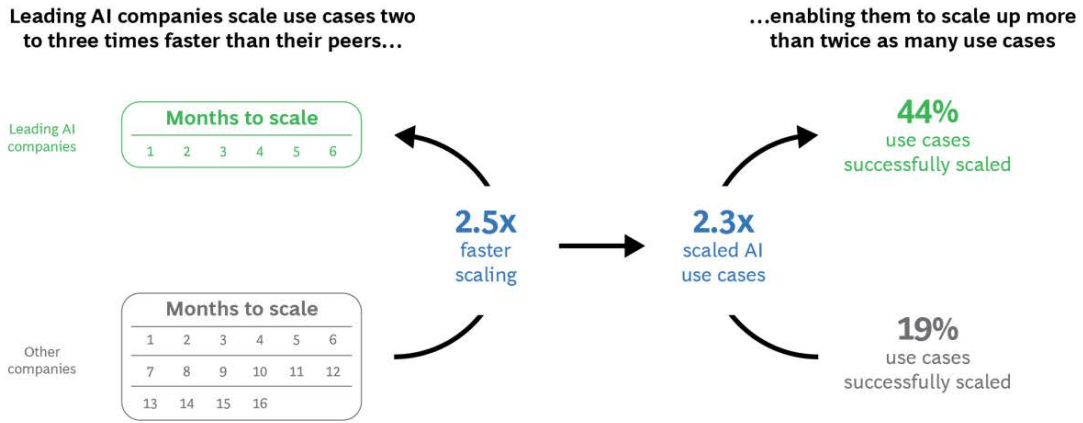
Community–focused

Leveraging international best practices

Work directly with policy makers and regulators

Why organizations care

Exhibit 3 - Leaders Scale More AI Use Cases More Quickly

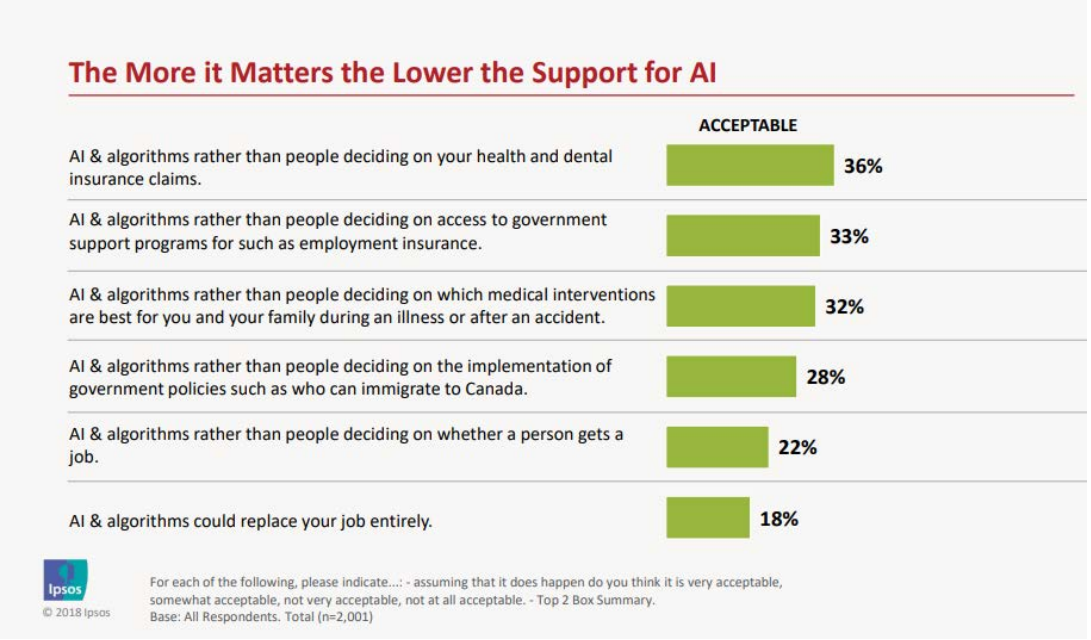


Source: BCG Digital Acceleration Index global study, 2022.

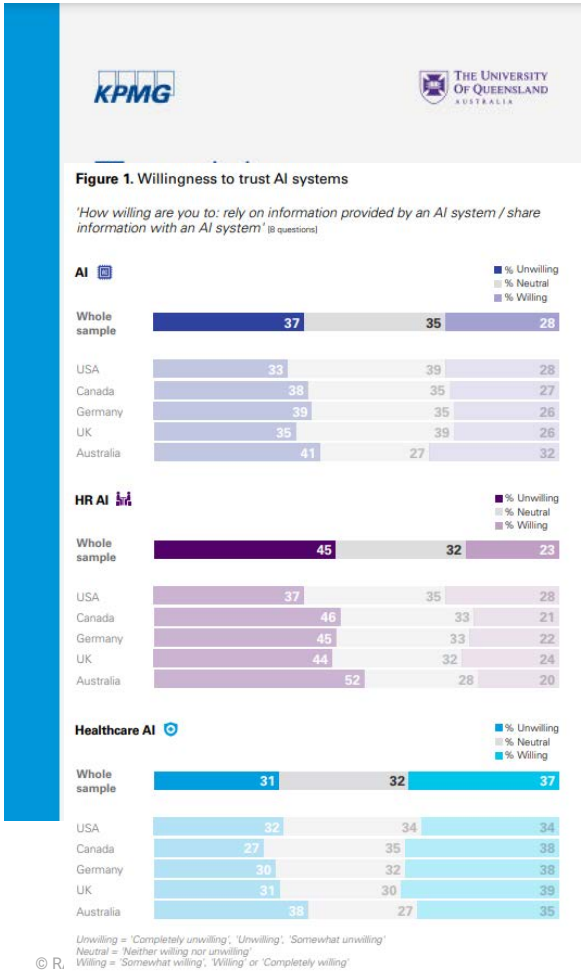
Disproportionate gains from getting AI right

BCG

But organizations know that public trust in AI is low

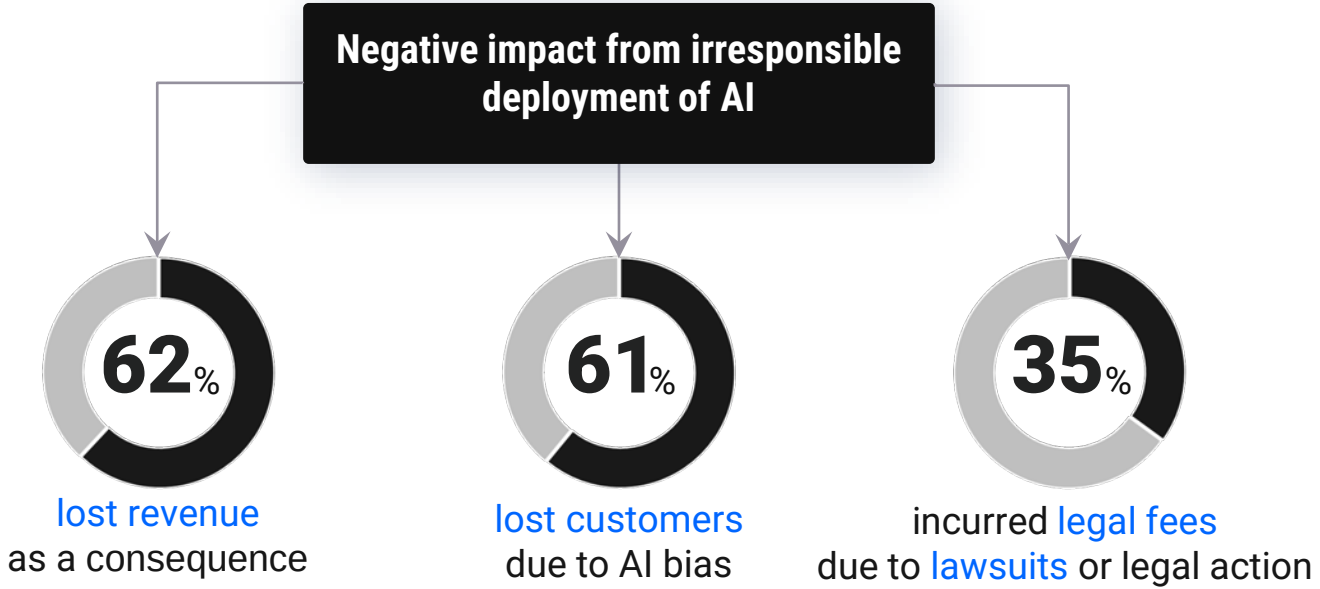


Tech for Good: A Canadian Perspective, Institut de Publique Sondage d'Opinion Secteur (IPSOS)



Trust in Artificial Intelligence, KPMG and the University of Queensland

Why responsible AI?



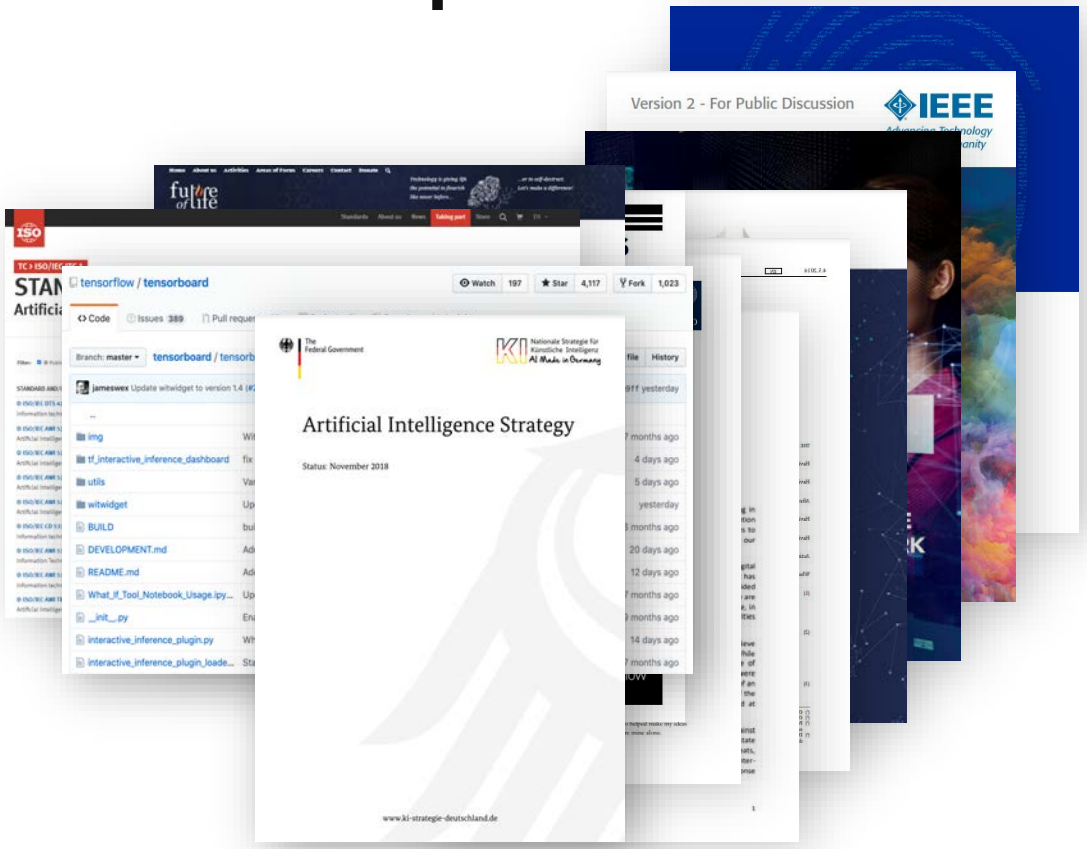
https://www.informationweek.com/big-data/the-cost-of-ai-bias-lower-revenue-lost-customers

Standards and Best Practices Landscape

There has been a significant demand from companies, academics, and governments for understanding RAI

Reports, research, principles, documents, and tools have emerged in response

While lots of good advice is readily available, it is difficult to navigate **what to do** and **how to implement**



AI Laws, Frameworks, and Standards

Emerging AI regulations generally expect AI systems to be:
Fair, explainable, accountable, robust, privacy preserving, protective of customers, and subject to effective human oversight

Leading efforts include EU's proposed AI Act, Canada's proposed AI and Data Act, NIST AI RMF, and ISO 42001 (AIMS)

Convergence around a risk–based approach, **Divergence** in the details

2023: Two Leading Auditable, Voluntary Standards Level

- **NIST AI Risk Management Framework, Jan 2023**
- **Important because**: US Gov aligns with NIST, and private sector follows its lead
- Q80 pieces of generic, industry-agnostic guidance
  - Eg “M3.1: Potential benefits of intended AI system functionality and performance are examined and documented.”
- Benefits:
  - Common vocabulary; plugs in w NIST Cyber and Privacy RMFs
  - Portal with Implementation profiles (use case eg. Fair housing, or temporal, eg. Current state medium size bank)

- **ISO 42001: AI Management Systems, ~Oct 2023**
- **Important because:** it can help show compliance with EU AI Act
- ~45 industry agnostic controls
- Significantly more detailed than NIST AI RMF
- Benefits:
  - Common vocabulary; plugs in w ISO 27001 (leading InfoSec standard)
  - More global (incl. non-Western), though lots of global companies default to NIST
  - Builds ‘documentation muscle’

Understanding the NIST AI RMF, Reva Schwartz (NIST), Ashley Casovan, Var Shankar (RAI Inst)

- describes NIST AI RMF and its relationship to ISO standards, legislation, and best practices

RAI Institute Implementation Framework

FOUNDATIONS

OECD · IEEE · European Commission · ISO · scc ccn · bsi. · VDE

RAI Institute Implementation Dimensions

1 Valid and Reliable
- 1.1 Data Relevance and Representativeness
- 1.2 Human–in–the–Loop
- 1.3 Guiding Policy Document/Strategy

2 Explainable and Interpretable
- 2.1 Communication About the Outcome
- 2.2 Notification

3 Accountable and Transparent
- 3.1 Team Training
- 3.2 Data Quality and Fit–For–Purpose

Aligned with NIST

4 Privacy–Enhanced
- 4.1 Transparency to Operators and End–User
- 4.2 Privacy Protection

5 Fair
- 5.1 Bias Impacts
- 5.2 Bias Training and Testing

6 Safe
- 6.1 Contingency Planning

7 Secure and Resilient
- 7.1 System Acceptance Test Performed

Community of Experts

- Industry experts
- Policy makers
- Academics
- Others

Key Use Cases

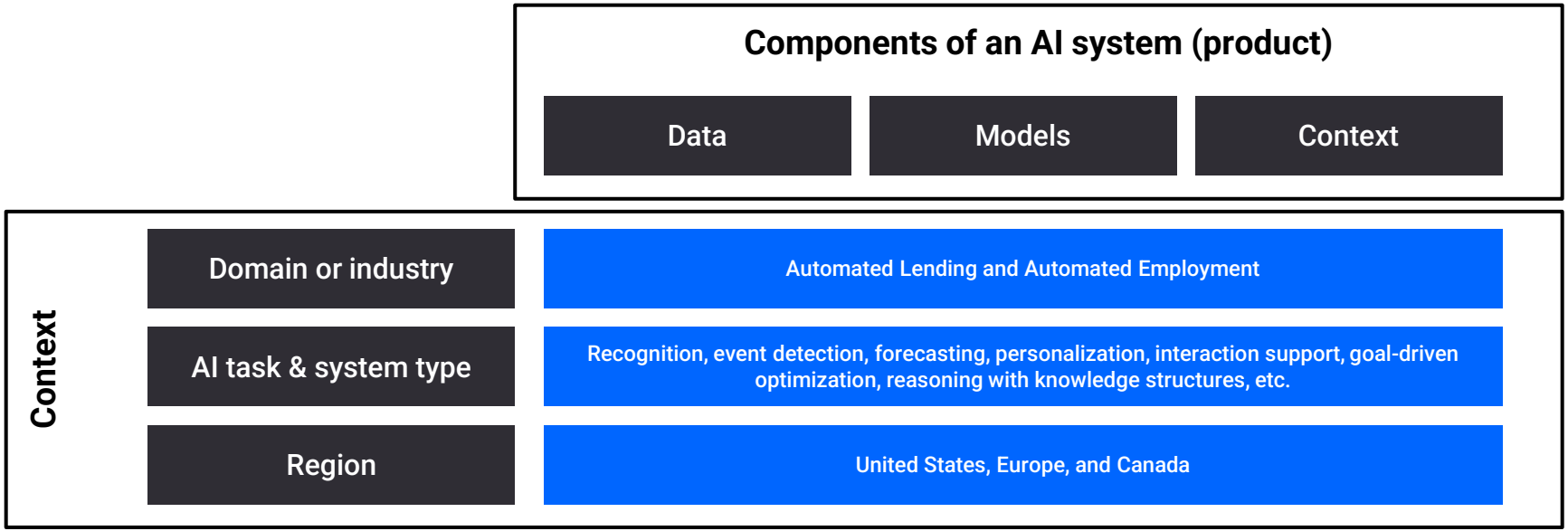
- Health care
- Human resources
- Procurement
- Financial services

AI Governance Ecosystem

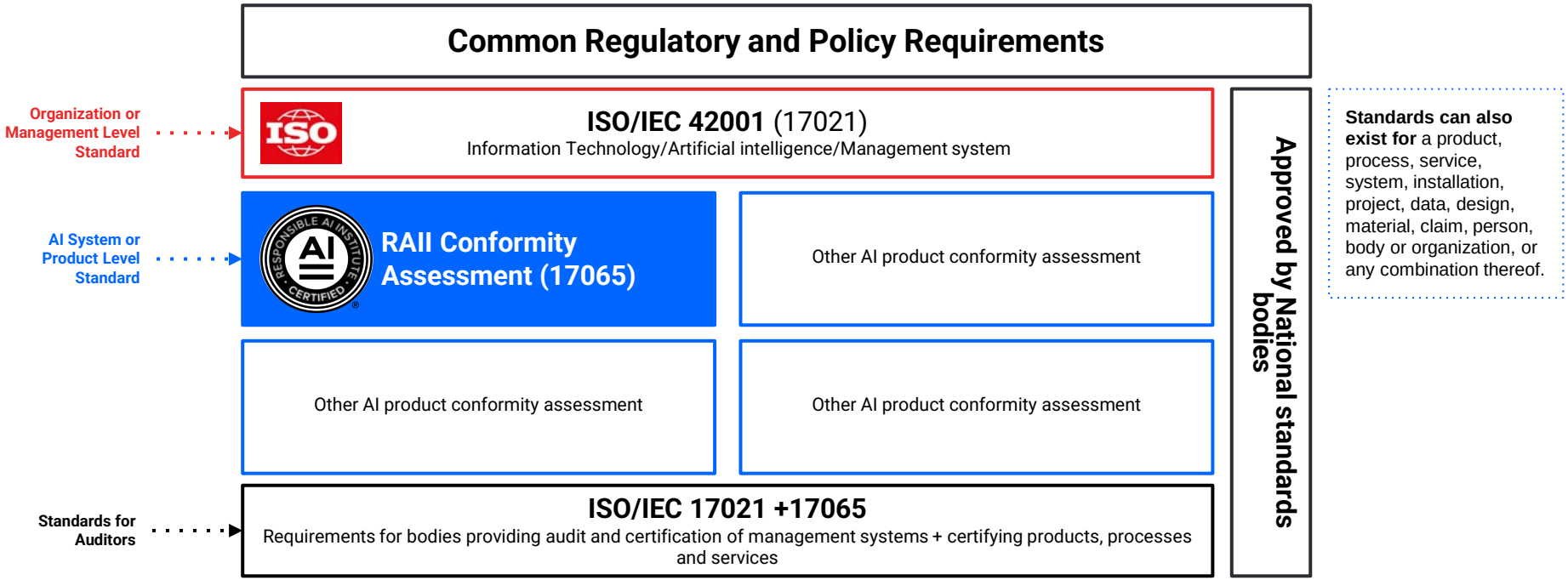
Drivers
Market/ Business
Consumer Demand

	Examples
Direction Setting Laws/ Acts	Draft EU AI Act (EU Commission) Bill C-27 - AIDA (Government of Canada) China Law on Management of Recommendation Systems (Government of China)
Sectoral or scoped rules Regulations	Road Traffic Act (Federal Ministry for Digital and Transport) National Traffic and Motor Vehicle Safety Act (US Department of Transport)
Standards	Automated Driving Taxonomy (SEA ISO J3016) Standard for the evaluation of autonomous products (ANSI/UL 4600) Road vehicles functional safety (ISO 26262)
Industry Best Practices	Society of Automobile Engineers (SAE)

Scope of a Product Conformity Assessment



Responsible AI Standards and Conformity Assessments



Demonstrating Commitment to Responsible AI

1st SELF-ATTESTATION OF CONFORMITY TO THE RESPONSIBLE AI INSTITUTE CERTIFICATION FRAMEWORK

2nd INDEPENDENT ATTESTATION OF CONFORMITY TO THE RESPONSIBLE AI INSTITUTE CERTIFICATION FRAMEWORK

RESPONSIBLE AI INSTITUTE · CERTIFIED ·

Self-Assessment
For low risk systems

Evaluate

Internal assessment of AI – enabled systems using the RAII Conformity Assessment

Second Party
For medium risk systems

Validate

Second party assessment of AI – enabled systems using the RAII Conformity Assessment via a workshop and second party review

Third Party
For high risk systems

Certify

Accredited independent audit of AI–enabled system using the RAII Conformity Assessments leading to RAII Certification

Methodology for Developing a Conformity Assessment

01 Convene group of relevant subject matter experts

Convene a diverse group of relevant stakeholders, including impacted individuals, AI developers and deployers, AI users, policy makers, researchers, and standards organizations to complete a landscape review of current issues associated with SLAs

02 Impact Identification

Complete an AI impact assessment (AIA) with relevant subject matter experts to identify harms and risks related to the AI system. Output of AIA is an impact mapping, including desired objectives and mitigation measures

03 Establish controls, test, and validate

Based on objectives and mitigation measures identified by subject matter experts in step 2, identify, test, and validate conformity assessment controls to be used by auditors. Once completed, conformity assessments are approved by national accreditation bodies

04 Accredit and Audit

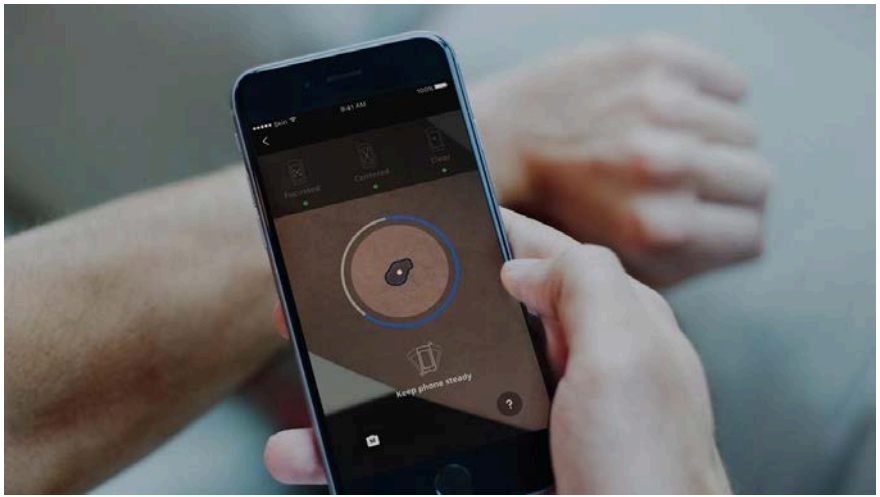
Once conformity assessment development is completed, national accreditation bodies review. Pilot conformity assessment through an audit process with two AI systems

Case Study at the System Level

Conformity Assessment for AI System

- Organizational maturity is necessary but not sufficient to focus on the use case

- **Use case: using smartphones to detect skin disease**
- Basis is Responsible AI Implementation Framework
- Add on FDA guidance and Clear Derm Guidelines on Image Based AI Diagnosis (from a group of leading dermatologists)
- Co-developing requirements with Memorial Sloan Kettering, Data Nutrition Project, IBM, Partnership on AI, Skinopathy

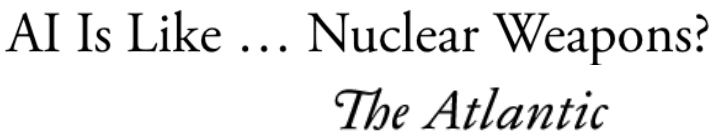


2/3 AI Regulation and Nuclear Regulation

Nuclear shows that we maintained democratic control of a disruptive technology in the past

- After Manhattan Project, there was pessimism about maintaining democratic control over a destructive new technology (eg Oppenheimer film)
- Yet, efforts to maintain democratic control have generally been successful
- **Government-to-government contacts**
- **Power plant monitoring** (eg monitoring environmental discharge required for nuclear power plants)
- **Controlling experts' access to knowledge and facilities**
- **Global monitoring** (eg International Monitoring System to detect nuclear tests)

Let’s acknowledge that AI is different



- AI is potentially dangerous and we are in a global arms race, so many have analogized to nuclear technology
- Since frontier models are expensive and require expertise, some have suggested a a regulatory regime similar to nuclear (ie a few, heavily regulated players)
- But generative AI models can be leaked or be open source – so anybody can propagate them immediately

What can AI Regulators Learn from Nuclear?

- Significant government attention, funding, and international collaboration (we're just starting to see serious attention and financial commitments)
- Dr. Heidy Khlaaf – Traceability of every component, rigorous documentation, and taking harms seriously (nuclear is seen as an existential threat, AI often seen as profit center)
- Clear objectives, common vocabularies & taxonomies, and ecosystem roles & responsibilities

3/3 CWG White Paper on AI Certification

What is the Certification Working Group?

- 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
- CWG is a multinational, interdisciplinary group of experts with academic, government, NGO, and corporate backgrounds in emerging technologies, law and policy, governance, evaluation, engineering, audit, standards, and certification
- Chair: Craig Shank

CWG White Paper

- **Unlocking the Power of AI – Steps for Effective Certification to Help Drive Innovation and Trust**
- Based on inputs from research, small group conversations, and interviews since 2021
- Early draft available at https://tinyurl.com/cwgdraft

CWG White Paper - Key Insights

- 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
- Regulatory regimes for AI – notably the EU AI Act – rely upon the development of well-developed AI certification ecosystem
- Developing a market-based AI certification ecosystem requires investments by key participants—like developers and certifiers
- Market signals to these key participants can help foster the development of the AI certification ecosystem

Gaps in AI Certification Ecosystem

- **Clarity** about what we are solving for – how to prioritize safety, trustworthiness, efficiency, profit, and other objectives in various contexts?

- **Reference architecture** outlining roles and responsibilities, such as whom an auditor or certification body should look to verify a given claim about a given implementation

- **Others:** Limited investment in certification ecosystem due to limited demand signals, limited transparency about how companies in the ethical AI “advisory” space are conducting their evaluations

Recommendations for Government

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

- **Support market development and demand signals for AI certification, including through regulation and procurement.**

Recommendations for All Stakeholders

- **Invest time and resources to get the foundations in place,** e.g. by clarifying what frameworks can be used for conformity assessment.

- **Collaborate to develop an effective AI reference architecture for policy and accountability,** enabling clarifications of roles, responsibilities (including shared responsibilities), exchange of documentation between suppliers at different points in the AI ecosystem necessary to deliver specific implementation, etc.

- **Move quickly to advance the state of the art from these foundations, e.g. by** developing a focused research agenda to support continued advancement in AI verification and validation tools to improve certification

Thank you

RAI Institute
www.responsible.ai
Contact Var Shankar: var@responsible.ai or LinkedIn