The standardization strategy on trustworthiness of AI in Japan

from the view-point of R&D and standardization

Dec 2nd 2021

Roy Sugimura

Supervisory Innovation Coordinator
Research Promotion Division for Artificial Intelligence of Information Technology and Human Factors,
National Institute of Advanced Industrial Science and Technology, Japan

What is “Trustworthiness”?

Definition of Trustworthiness in ISO/IEC TR 24028:2020(en)3.42

trustworthiness

ability to meet stakeholders’ expectations in a verifiable way

Note 1 to entry: Depending on the context or sector, and also on the specific product or service, data, and technology used, different characteristics apply and need verification to ensure stakeholders expectations are met.

Note 2 to entry: Characteristics of trustworthiness include, for instance, reliability, availability, resilience, security, privacy, safety, accountability, transparency, integrity, authenticity, quality, usability.

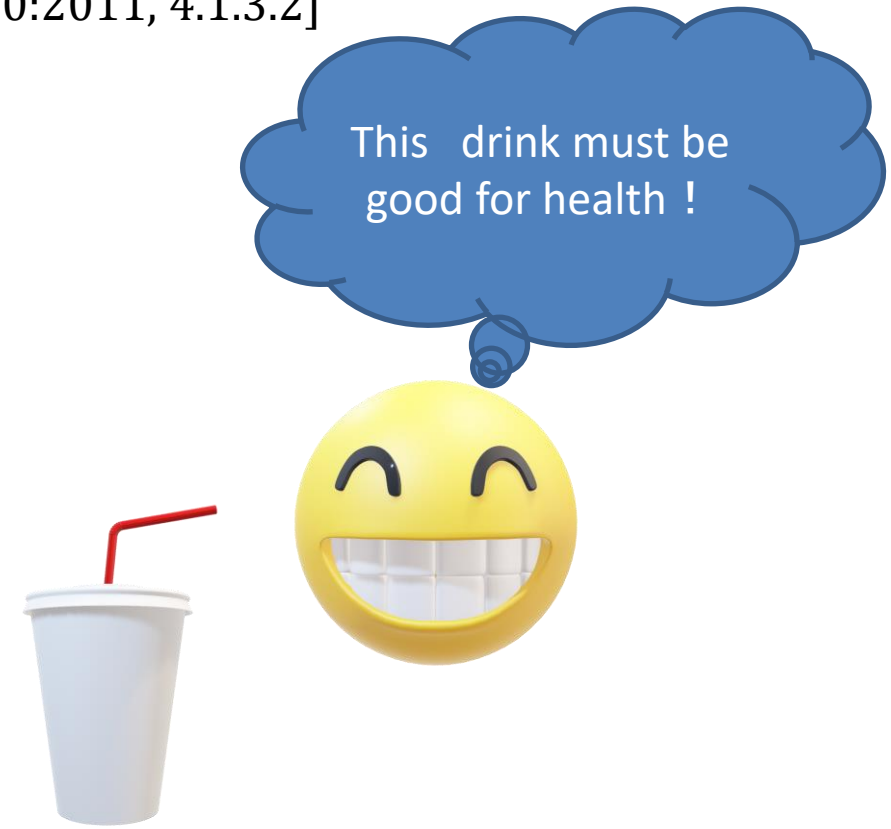
Note 3 to entry: Trustworthiness is an attribute that can be applied to services, products, technology, data and information as well as, in the context of governance, to organizations.

ISO/IEC TR 24028:2020(en) 3.42

Artificial intelligence (AI) — Overview of trustworthiness in artificial intelligence

(cf) Any Difference between trust and trustworthiness?

trust
degree to which a user or other stakeholder has confidence that a product or system will behave as intended
[SOURCE: ISO/IEC 25010:2011, 4.1.3.2]



Any leading direction for trustworthiness?

From “Integrated Innovaiton Strategy 2021

Cabinet Decision, Published by Cabinet Office (in Japanese)
https://www8.cao.go.jp/cstp/tougosenryaku/2021.html

Chapter 2,
Science, Technology and Innovation Policy for Society 5.0

1. Transformation into a sustainable and resilient society that ensures the safety and security of the people

(1) Creation of new value through the fusion of cyberspace and physical space

[Target]

・Complete the "Data Strategy" and transform cyberspace and physical space into a society that creates a dynamic virtuous circle, so that anyone, anywhere, anytime can create new value by utilizing data and AI with **confidence**. (pp 24)

<snip>

③ Establishment of a **reliable** data distribution environment including data governance rules (pp 27)

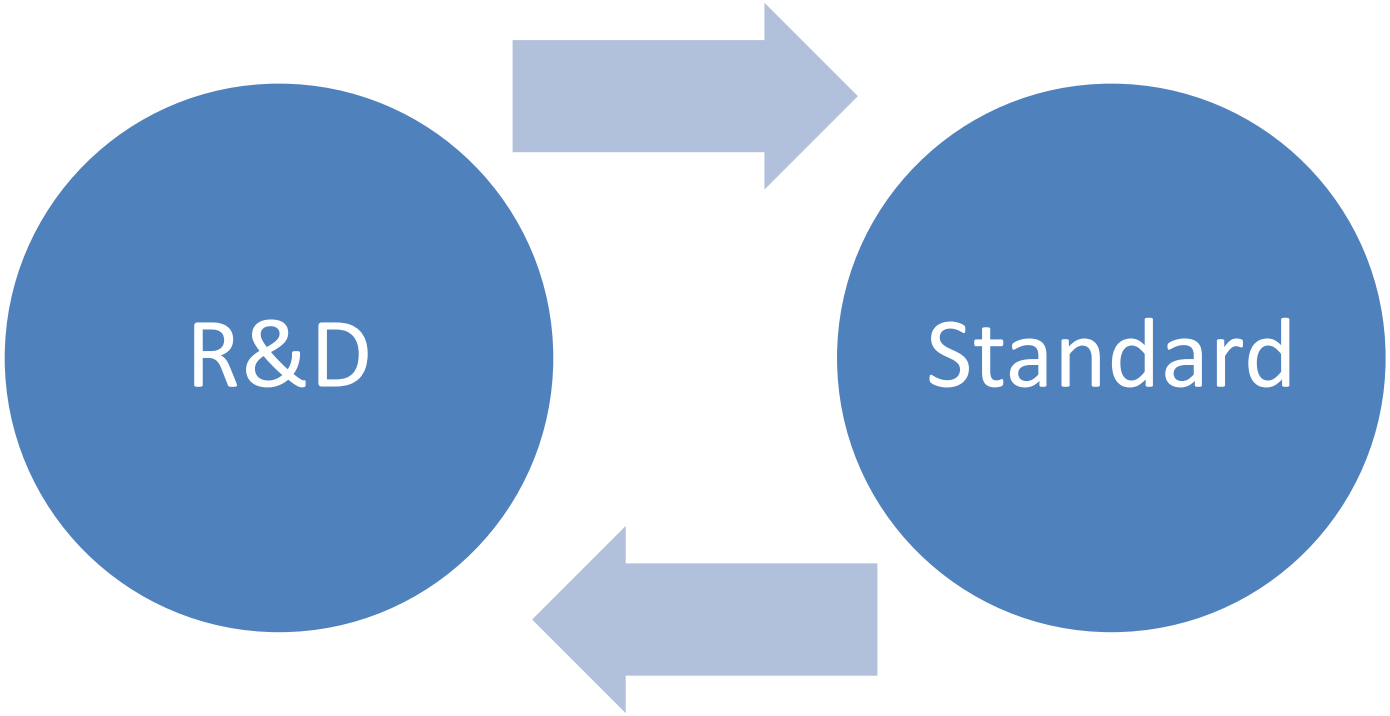
④ Development and R & D of next-generation infrastructure and technologies for data and AI applications in response to the digital society

“To realize a next-generation social infrastructure suitable for the use of data and AI in terms of power saving, high **reliability**, and low latency, which is laid out in a network throughout the country.” (pp.28)

(continued to the next page)

4. Promotion of sectoral strategies through public-private partnerships
(Basic Technologies to Be Strategically Addressed)
(1) AI technology
<snip>
The strategy will be continuously reviewed based on the progress of the strategy and progress in the implementation of AI in society. This will include next-generation machine learning algorithms based on the principles of deep learning, advanced natural language processing such as simultaneous interpretation, and highly **reliable** AI, which is important for applications in the medical and manufacturing fields. The strategy will also be promoted so that each and every citizen can realize the specific benefits of AI.

Standardization & R&D



Progress of "AI Strategy 2019", review of "AI Strategy 2021" and formulation of new AI strategy by Cabinet Office - Office for the Promotion of Science, Technology and Innovation

II. Building Foundations for the Future : Educational Reform and Reconstruction of the R & D System
<snip>
II-2 Rebuilding the R & D System
<snip>
(2) Launch of core research programs : Promotion of basic and integrated research and development
<Specific Target>
Strategically promote the research and development of basic and integrated AI technologies (AI Core), which are important for achieving major goals, by organizing them into the following four areas

1. Basic Theories and Technologies of AI
2. Device and Architecture for AI
3. **Trusted** Quality AI
4. System Components of AI

R&D

III. Building the foundations of industry and society
III-2 Development of Data Infrastructure
(2) Trust and Security
<Specific Target 1>
Establishment and development of a trust data collaboration platform capable of international mutual authentication with the United States, Europe, etc.
・Promotion of international standardization related to AI life cycle and **AI quality assurance**, including ensuring **data quality** (FY 2021)

Key words around Trustworthiness in summary

The Use of Data and AI in terms of high reliability

The strategy will be continuously reviewed based on the progress of the strategy and progress in the implementation of AI in society. This will include next-generation machine learning algorithms based on the principles of deep learning, advanced natural language processing such as simultaneous interpretation, and highly **reliable** AI, which is important for applications in the medical and manufacturing fields. The strategy will also be promoted so that each and every citizen can realize the specific benefits of AI.

R&D

Strategically promote the R&D of basic and integrated AI technologies (AI Core), which are important for achieving major goals, by organizing them into the following four areas

1. Basic Theories and Technologies of AI
2. Device and Architecture for AI
3. **Trusted** Quality AI
4. System Components of AI

Standard

Establishment and development of a trust data collaboration platform capable of international mutual authentication with the United States, Europe, etc.

・Promotion of international **standardization** related to AI life cycle and **AI quality assurance**, including ensuring **data quality** (FY 2021)

R&D

AIRC, AIP, and AIS

- AIRC/ AIST was established in May 2015 to be the largest AI research center in Japan for promoting large-scale AI research with PPP.
- Cooperating with RIKEN and NICT, AIRC/AIST accelerates AI R&D and deployment with industries and overseas research institutes.

Organization	established	Focused Research
AIRC: Artificial Intelligence Research Center AIST/ METI	May 2015	• R&D for deployment of AI (productivity, medical &welfare, transport, etc.) • Strategic AI research, AI platform and AI infrastructure
AIP: Center for Advanced Intelligence Project Riken /MEXT	April 2016	• Theoretical research on machine learning • Goal oriented basic research • Social research on AI
AIS: AI Science R&D Promotion Center NICT/ MIC	April 2017	• Brain Research • Communication research, including automatic translation systems

R&D

AIST R&D activities for AI

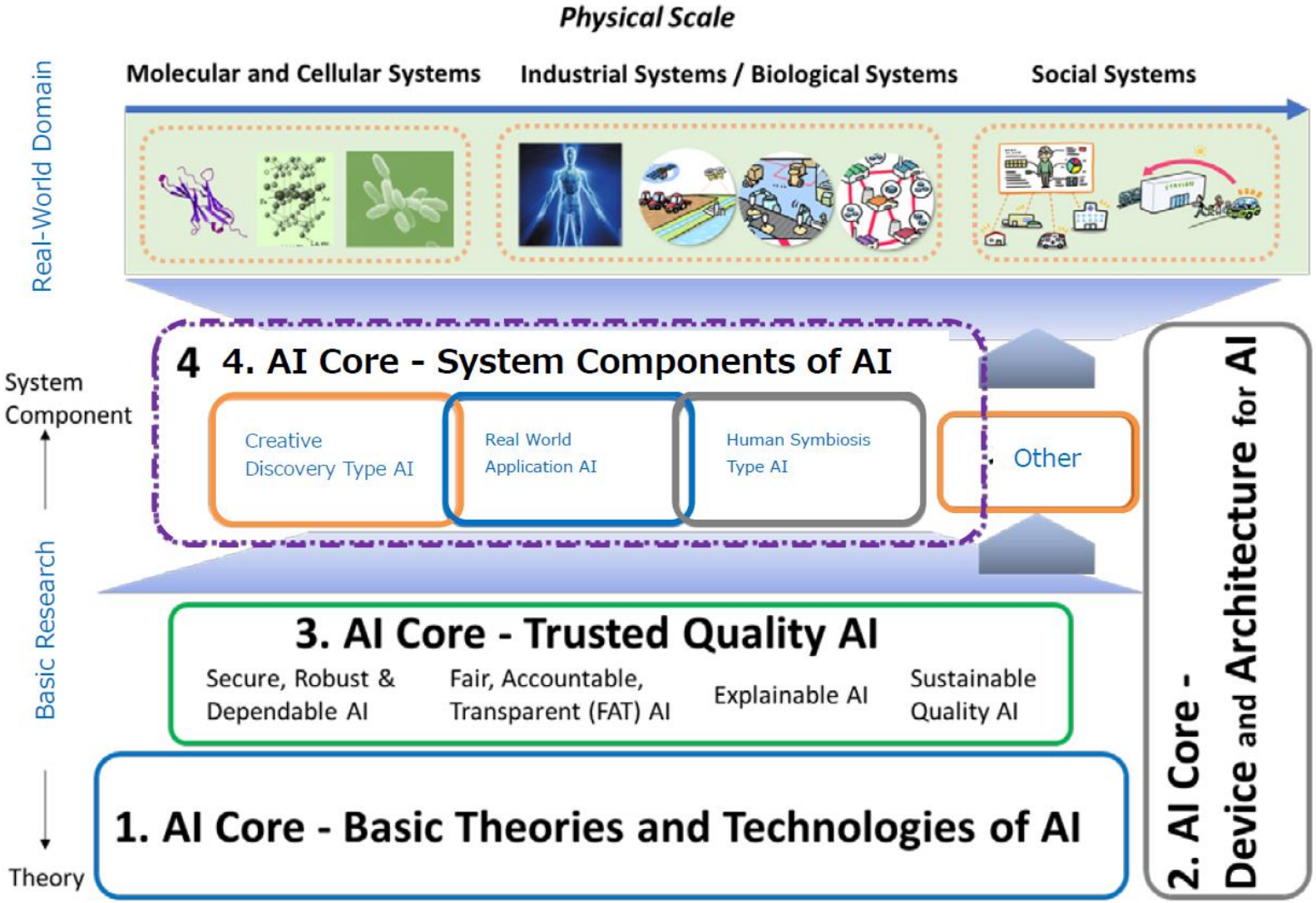


Figure: Overall Structure of AI R&D

R&D

Our three pillars for development on Basic AI technology in AIST

＜Easy-to-construct AI＞

- From **AutoML to AutoAI**
- Scalable AI, Composition of AI Modules, **Transfer Learning**
- Standardization,Connectivity

＜AI which cooperates with Human＞

- **Explainability of AI**
- Integration of ML and knowledge
- Interaction and Co-evolution
- Acquisition of Tacit Knowledge and Skills

＜Trustworthy, Reliable AI＞

- Verification and validation
- **Quality Assessment of AI systems**
- **Guarantee of AI systems**, Constraints

Modified the resource at
https://www.airc.aist.go.jp/info_details/docs/180523/ai_strategy180523.pdf

R&D

Network of Partners as a "Center"

【Research Labs.】

- National Cancer Center Japan
- National laboratories of MLIT, MAFF etc

【RIKEN & NICT】

- Joint researches and symposium
- Cooperation in use of HPC

【Industries】

- Joint laboratories w/ NEC etc.
- Cooperative researches w/ industries : ～50 (Total~170)
- AI Technology Consortium: ～170 firms participated

AIST/AI Research Center

【Domestic Univ.】

- Network with university researchers : more than 80 (～30 domestic universities , national laboratories, basic private research institutes)
- Students attended : ～80

【Overseas Labs.】

- EU/ US's core univ./ institutes (U of Manchester, DFKI, CMU, UCSD etc.)
- Asian univs. /institutes networks (SG, Taiwan, Thailand, India)
- Foreign FT researchers : ～30%, foreign researchers & students: ～80 (from 20 countries)

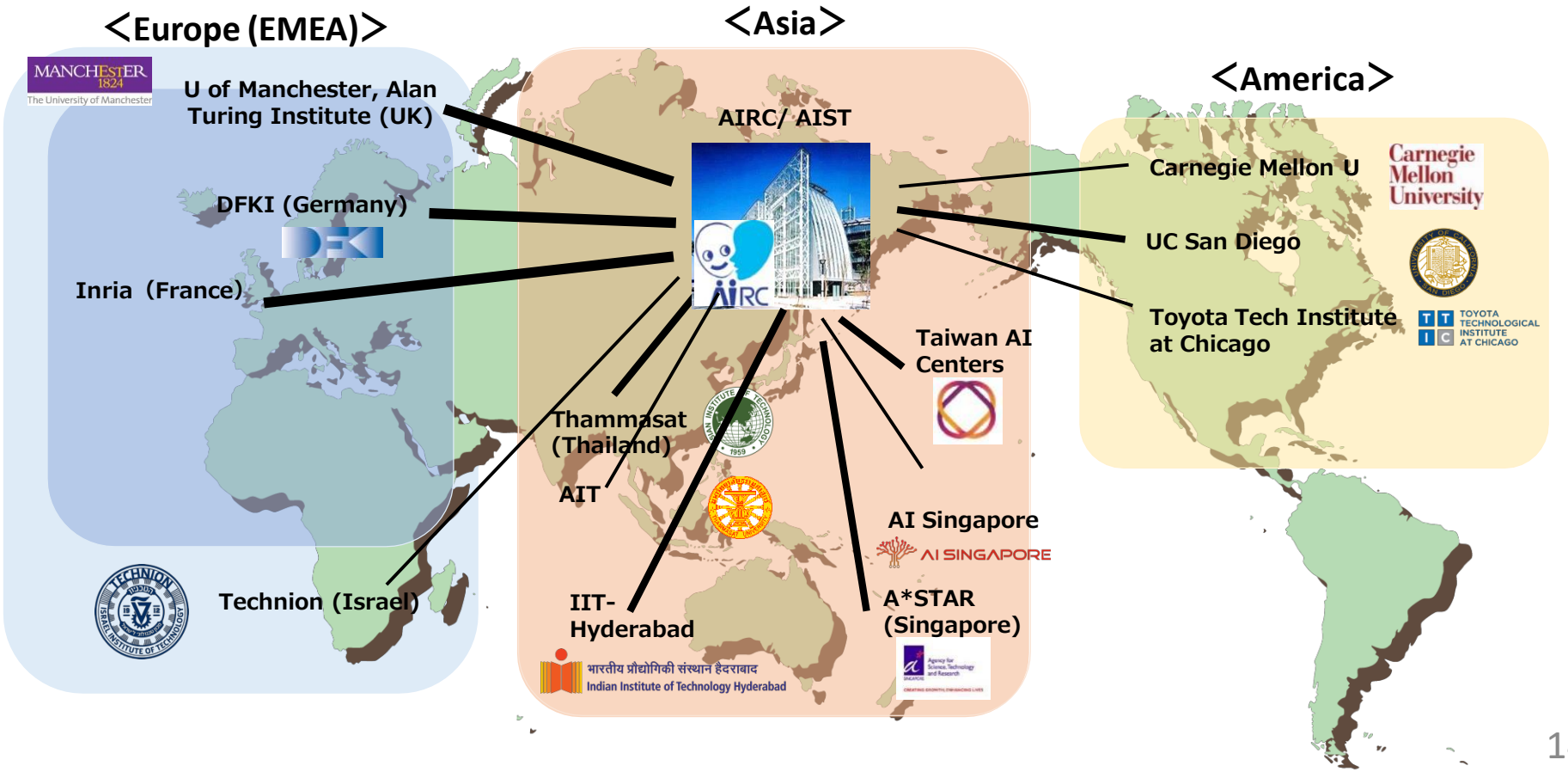
【 Outreach - Diffusion & HR dev.】

- AI seminar every 1-2 months. Lecturers at various kinds of seminars.
- Cooperation w/ universities for HR resources development

R&D

Global Network of AIRC in AIST

- Many countries in the world are now trying to establish their own AI strategies and to create “AI centers” then to collaborate globally - especially with in mind of tech giants in the US and China.
- AIRC has been creating close relationship with EU and US core AI institutes/ universities, and then with Asian institutes/ universities to facilitate Asian AI research networks.



Standard

Machine Learning Quality Management Guideline by AIST

National Institute of
Advanced Industrial Science and Technology

Rev. 1.0.1.0037-e19 (2021/03/09)

Machine Learning Quality Management Guideline

1st English Edition

February 12, 2021
(Japanese: June 30, 2020)

Minor Update 1: March 9, 2021

Technical Report CPSEC-TR-2020002
Cyber Physical Security Research Center
National Institute of Advanced Industrial Science and Technology (AIST)

Standard

Contribution to Standardization

Collaboration with our respected organizations in SC42.
Collaboration with our respected regional organizations in EU, USA, and Asia directly and also through high-level layers in METI.

Providing experts for standardization.

Conveners	Reference	Document title	Developing committee	Project leader	PL国
Prof. Harada	DIS 38507	IT — Governance of IT — Governance implications of the use of AI by organizations	SC42/JWG 1	Peter Brown	UK
	AWI 5259-1	Data quality for analytics and ML — Part 1: Overview, terminology, and examples	SC42/WG 2	Suwook Ha	Korea
	AWI 5259-2	Data quality for analytics and ML — Part 2: Data quality measures	SC42/WG 2	Kyoung-Sook Kim	Japan
	AWI 5259-3	Data quality for analytics and ML — Part 3: Data quality management requirements and guidelines	SC42/WG 2	Matthis Eicher	Germany
Mr. Enomoto	AWI 5259-4	Data quality for analytics and ML — Part 4: Data quality process framework	SC42/WG 2	Wanzhong Ma	China
	AWI TR 5469	AI — Functional safety and AI systems	SC42/WG 3	Takashi Egawa	Japan
	AWI 5338	IT — AI — AI system life cycle processes	SC42/WG 4	Yuchang Cheng	Japan
	AWI 5339	IT — AI — Guidelines for AI applications	SC42/WG 4	Shrikant Bhat	India
Dr. Maruyama	TR 24030:2021	IT — AI — Use cases	SC42/WG 4	Yuchang Cheng	Japan

Trustworthy AI
Strategic Hypothesis

- High Quality AI
  - Collaboration with global R&D and Standardization partners : AIST, JISC
  - Several activities around International Standardization initiated
    - Machine Learning Quality Management Guideline : AIST
    - Data related : ISO/IEC JTC 1/SC 42 WG 2
    - Functional Safety related: ISO/IEC JTC 1/SC 42 WG 3

Further Challenge

- AI which cooperates with Human : Human Machine Teaming
- Bias handling in AI algorithm and Data
- Multi-modal AI : Inductive and Deductive knowledge
- etc.

- Hypothesis generation under partial information
- Quick decision and action
- Monitoring and quick feed-back
- Reflection and redirection

with

Trusted Global Teamwork
Respect among members

Thank you for your attention