Implementation of AI Safety Evaluation in the Healthcare Sector

Japan AI Safety Institute (AISI) –
Business Demonstration WG
Healthcare SWG

AISI Japan
AI Safety Institute

Speaker Introduction

Two speakers from Ubie, Inc., the leader company of the Healthcare Sub-Working Group (SWG), will be presenting.

Mamu Inoue

Head of the Accelerator Department /Director of Government Affairs, Ubie, Inc.
Leader, WG4 Sub-WG-B, Japan Digital Health Alliance

Former officer at the Ministry of Internal Affairs and Communications with eight years of experience in digital and telecommunications policy.

Joined Ubie in 2022, leading the company’s public and govermental affairs team, including the development of the “Guide to the Utilization of Generative AI for Healthcare Providers”.

Masahiro Kazama

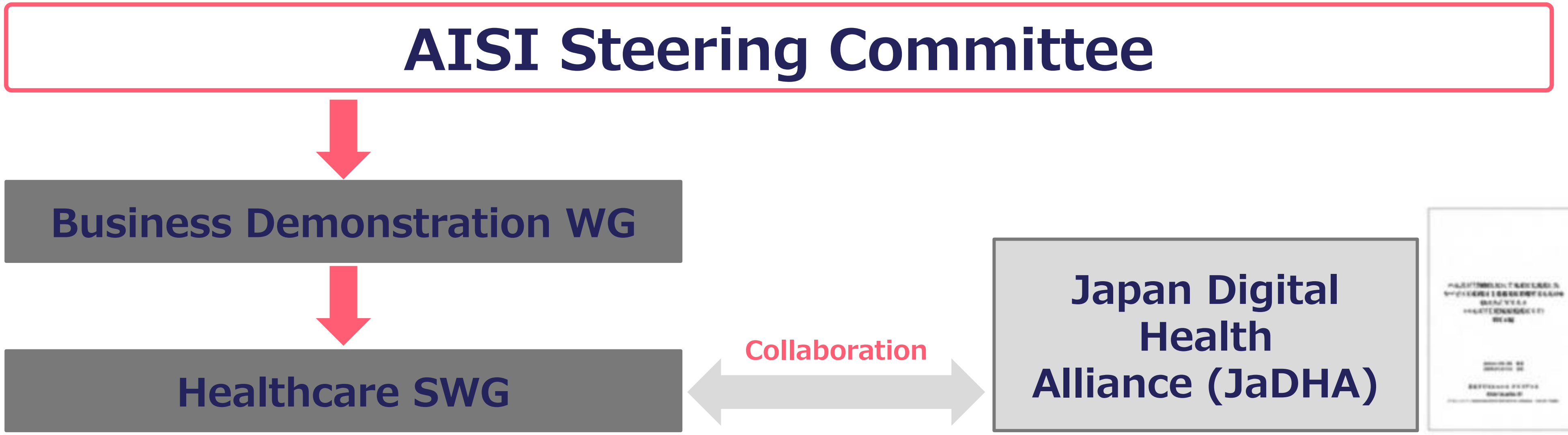
Chief AI Officer (CAIO), Ubie, Inc.
Adjunct Lecturer, Tsuda University
Visiting Researcher, National Institute for Japanese Language and Linguistics (NINJAL)

Graduated from the University of Tokyo Graduate School in 2015 and joined Recruit Holdings and Indeed; joined Ubie in 2020 to lead AI medical interview systems and now oversees generative AI initiatives.

Forbes 30 Under 30 Japan (2018)
Author of "Introduction to Recommender Systems in Practice" (O'Reilly Japan, 2022)

AISI Healthcare SWG

The Japan AI Safety Institute (AISI) established the Healthcare Sub-Working Group (SWG) to promote AI safety evaluation in the healthcare sector. The SWG consists of domain experts and works closely with industry associations to develop practical rules and guidance.



AISI Steering Committee

Business Demonstration WG

Healthcare SWG

Collaboration

Japan Digital Health Alliance (JaDHA)

Ubie (SWG Leader)
Awarefy, CMIC Holdings, MICIN, etc.

Collaboration with WG4 SubWG-B
(Working Group on Generative AI)

JaDHA is an industry association with over 100 member companies, including major pharmaceutical and medical device manufacturers, healthcare ventures, and ICT companies.

Why AI Safety Evaluation Matters in Healthcare

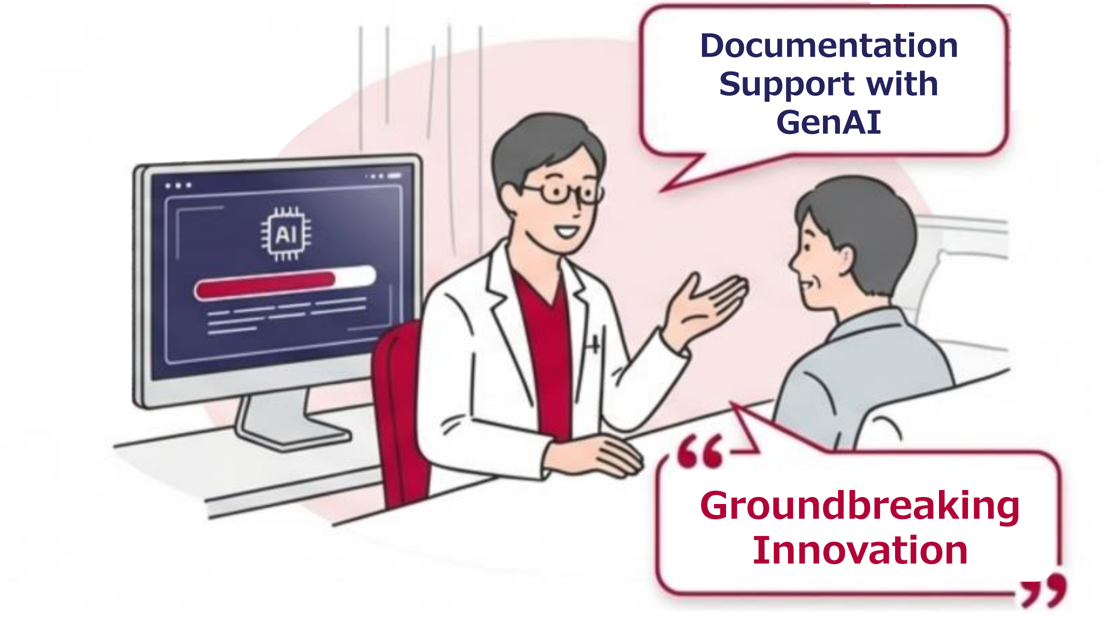
- Generative AI is a groundbreaking innovation that is expected to enhance productivity of healthcare professionals and enable more personalized care.
- However, the healthcare sector involves the handling of highly sensitive information and can impact patients’ life and health.
- Therefore, innovation must be balanced with building a safe and trustworthy environment for AI usage.

Innovation for the Future of Healthcare

Japan's healthcare system faces challenges such as work-style reform and increasing medical demand.

Among reasons for physicians' overtime work, "clerical duties (recordkeeping, report writing, etc.)" (59.8%) ranks second after "patient care".

Source: Special site for Work Style Reform for Doctors



Documentation Support with GenAI

"Groundbreaking Innovation"

Generative AI can help healthcare professionals focus on their core clinical work.

Scope of the First Year Activities

Target Users: AI providers

Organizations/businesses that develop products and services using foundation AI models or APIs.

Target Product: Non-SaMD

Products/services that are not classified as Software as a Medical Device (SaMD).

Target Technology: Text-generation AI (LLM)

Text-generation AI (LLMs) is currently the most widely used AI in healthcare.

Use case example: Discharge Summary Support

We aim to reduce physicians' workload by using generative AI to summarize clinical notes and generate discharge summary templates.

Input

EHR data
(medical records, test results, prescriptions, etc.)

Generation

Before

Manual work by physician

After

AI Automation

Summarization + Generation of Draft by GenAI

Output

Discharge Summary

Risk Scenarios & Evaluation Points: Discharge Summary Support

Prevention of Misinformation, Disinformation and Manipulation

Risk Examples) AI generates diagnoses or medication information not present in the medical record. Critical details (e.g., allergies) are omitted from the summary.

Hallucination control, Summarization accuracy

Privacy Protection

Risk Examples) Insufficient masking of patients’ personal information (name, diagnosis, etc.). External leakage of such information or inclusion in input/output.

Appropriateness of masking/redaction

Ensuring Security

Risk Examples) Network vulnerabilities causes leakage of hospital information via the generative AI system.

Security of the network environment

Robustness

Risk Examples) If outputs vary under identical conditions, reliability as a medical record is undermined.

Output consistency, Variance

Examine healthcare-specific scenarios based on the 10 evaluation items in AISI’s “Guide to Evaluation Perspectives on AI Safety”

Planned Deliverables: AI Safety Guide and Evaluation Checklist

Our findings will be consolidated into a practical “AI Safety Guide” that will be continuously updated. An “Evaluation Checklist” and example prompts will also be published to encourage adoption.

1. Background, objectives & target users:
Why AI safety matters in healthcare

2. Trends in AI safety for healthcare:
Technology, industry, and regulatory trends (Japan and abroad)

3. What to Evaluate:
Evaluation perspectives and risks in healthcare

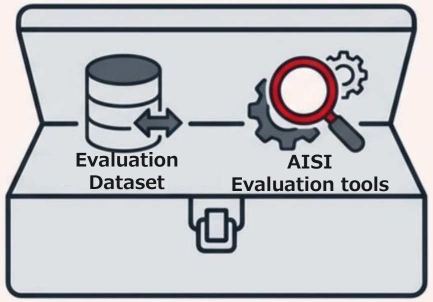
4. How to Implement:
Metrics, datasets, and evaluation tools

5. Outlook: Future directions

Evaluation checklist

Covers items to check at each stage of development—helping prevent overlooked risks and enabling systematic safety assurance.

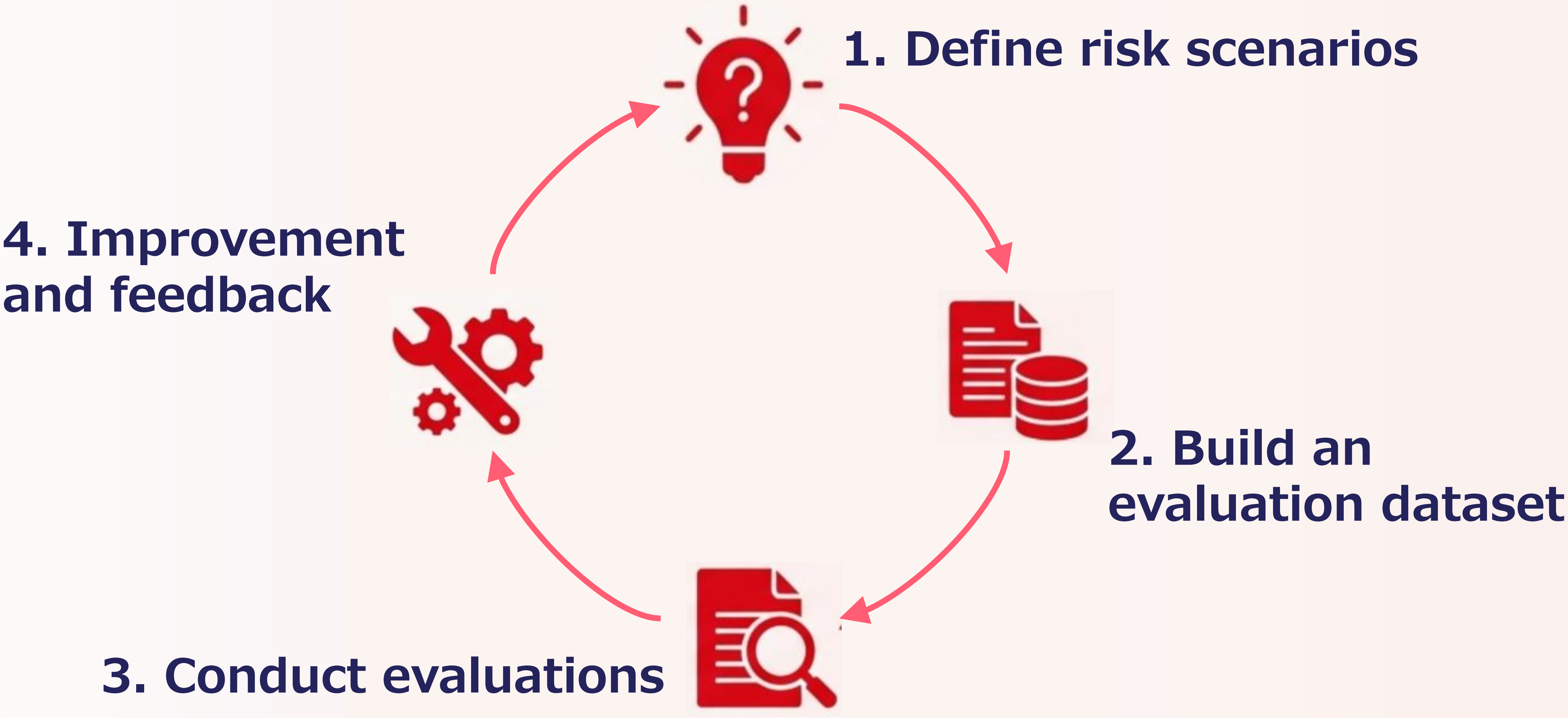
Qualitative & quantitative methods



In addition to qualitative evaluation using the checklist, explains how to conduct quantitative evaluations using evaluation datasets and AISI evaluation tools.

Embedding AI Safety into the Development Lifecycle

We are developing guidelines designed for easy integration to continuously verify and improve safety throughout the product development lifecycle



1. Define risk scenarios

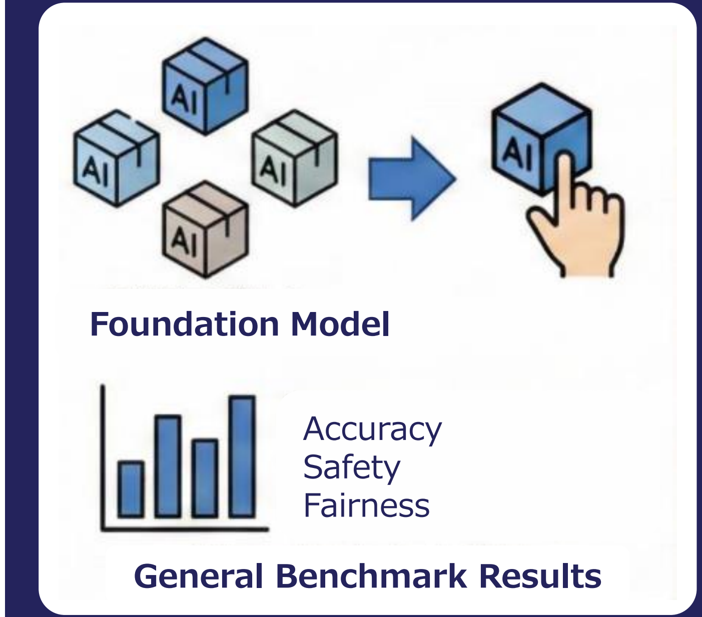
2. Build an evaluation dataset

3. Conduct evaluations

4. Improvement and feedback

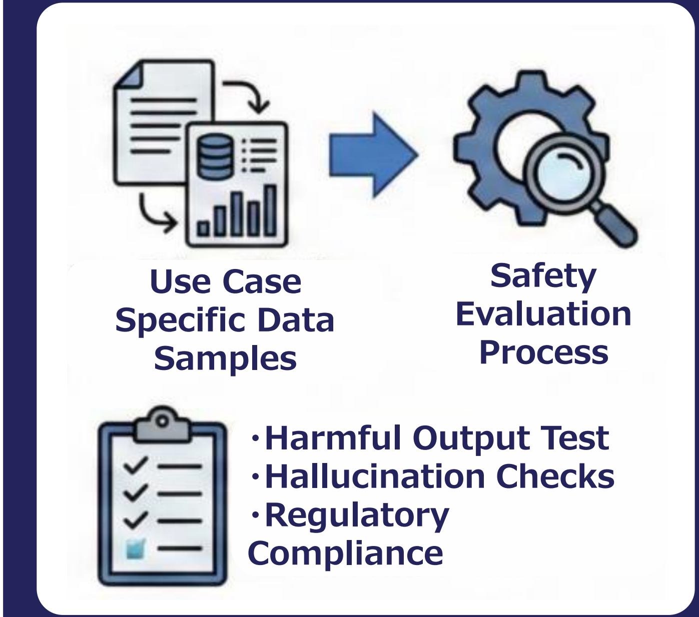
Evaluation and Risk Mitigation Process

① Model selection phase



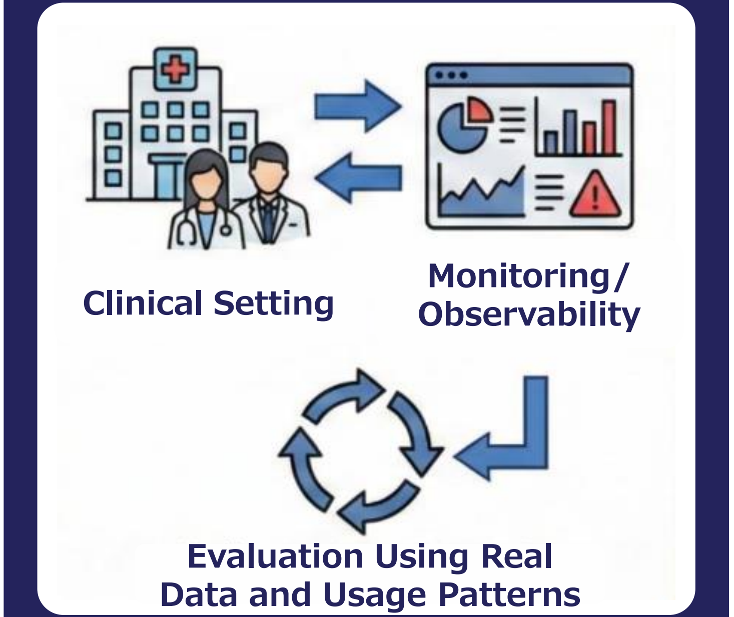
- Select a foundation model
- Select model based on general benchmark results

② Product development phase



- Create data samples tailored to your use case
- Conduct safety evaluations

③ Product operation (deployment) phase


Clinical Setting
Monitoring / Observability
Evaluation Using Real Data and Usage Patterns

- Monitor and ensure observability in real clinical settings
- Evaluate safety using real data and usage patterns

Use appropriate methods in each phase to continuously ensure safety

Next Steps: From Implementation to Standards

Short-term Initiatives (FY2025)

Foundation

- Select representative use cases for Non-SaMD
- Clarify the risk structure and design safety evaluation scenarios
- Develop and publish the first edition of the guide

Medium-term Initiatives (FY2026~2027)

Validation & Expansion

- Validate the effectiveness of the guide/tools with AI developers and providers (e.g., pilot evaluations at multiple medical institutions)
- Consider expanding the scope of target products & technology (SaMD, multimodal, etc.)

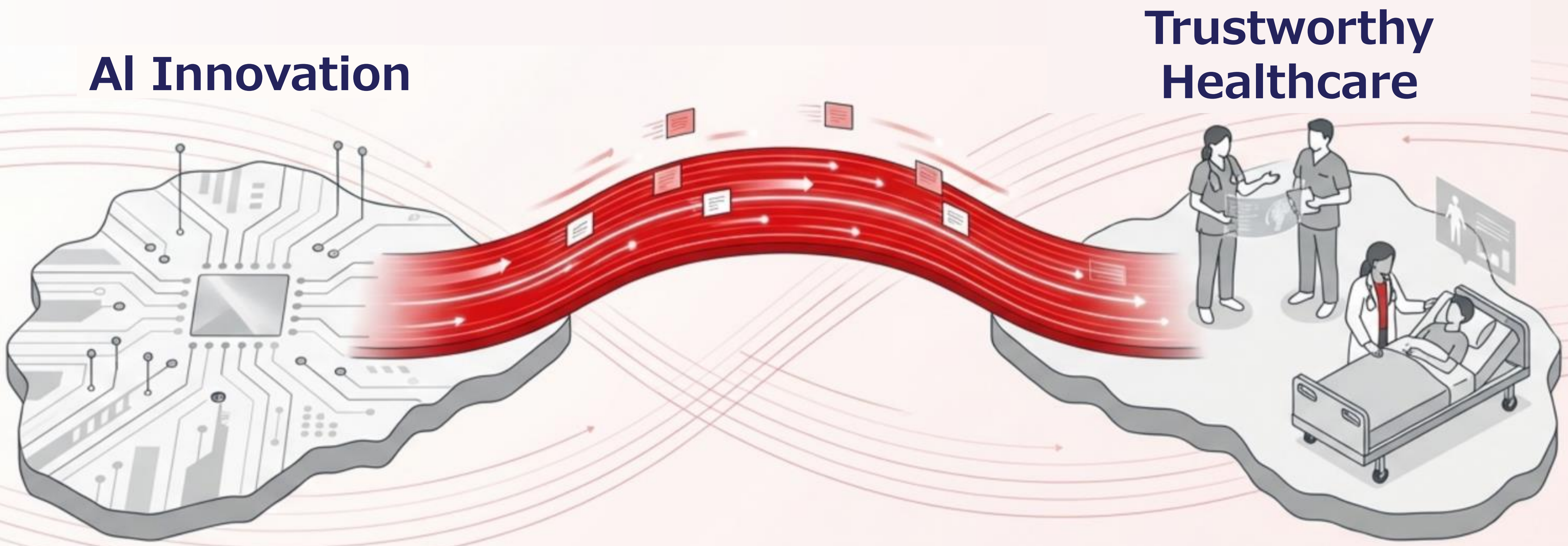
Long-term Initiatives (Future Vision)

Standardization & Dissemination

- Promote and drive adoption of the guide/tools across the healthcare industry
- Establish a mechanism for continuous updates in response to technological and societal changes

Coexistence of AI Trustworthiness and Innovation

AISI Japan AI Safety Institute



AI Safety is essential to fully unlock the potential of AI and enhance the quality of healthcare.
Our work is a concrete step towards that goal.

AISI

Japan AI Safety Institute