Guide to Evaluation Perspectives on AI Safety in the Healthcare Sector

— Toward the Realization of Trustworthy AI —

(Summary) ver1.0

Japan AI Safety Institute
Business Demonstration Working Group
Healthcare Sub-Working Group

April 3, 2026

AISI Japan AI Safety Institute

Structure of the Guide

AISI Japan AI Safety Institute

- Drawing on global and domestic trends toward the realization of trustworthy AI, this guide provides practical guidelines for AI Safety evaluation specifically tailored to the healthcare sector.
- This guide is designed with a focus on practicality, incorporating the evaluation perspectives of AISI* while also organizing evaluation items for each phase assuming actual operations.

*"Guide to Evaluation Perspectives on AI Safety" (AISI) https://aisi.go.jp/output/output_framework/guide_to_evaluation_perspective_on_ai_safety/

Structure of the Guide

1. **Background and Purpose of the Guide**
   - The importance of AI Safety in the healthcare sector and the scope of the guide (AI providers, Non-SaMD, text generation)
2. **AI Trends in the Medical & Healthcare Sector**
   - Technological trends, policy/industry trends, and market trends
3. **10 Evaluation Perspectives on AI Safety**
   - Overview of each perspective, potential risks, actual cases, and examples of evaluation items when applying the "Guide to Evaluation Perspectives on AI Safety" (AISI) to the healthcare sector
4. **Practical Implementation of AI Safety Evaluation in AI Product Development**
   - Key evaluation perspectives by phase, specific implementation methods, and a checklist of points to check
5. **Future Outlook**
   - Directions for the further development of the guide

Chapter 1: Background and Purpose of the Guide

AISI Japan AI Safety Institute

- The scope of this guide is on AI providers, Non-SaMD (AI products that are not classified as Software as a Medical Device), and text-generating AI (LLM). Both B2B (for healthcare professionals) and B2C (for consumers) use cases were considered.
- The intended audience includes executives, product managers, engineers, and legal professionals involved in the planning, development, operation, and risk assessment of healthcare AI services.

Scope of the Guide

Target Audience	**AI Providers** (Businesses that develop products or services using pre-trained generative AI models via APIs or other means)
Target Products	**Non-Software as a Medical Device (Non-SaMD)**
Target Generative AI	**LLMs (Text-based generative AI)** (Image-based and voice-based generative AI, etc., are out of scope of this guide)

Target Use Cases

Category	Example Use Cases
B2B (For Healthcare Professionals)	Document preparation support, information retrieval and summarization, support for medical record entry, support for creating patient information materials, medical literature search and summarization, etc.
B2C (For Consumers)	Health consultation chatbots, self-care support, mental health care apps, medication reminders, health management apps, etc.
Other	Support for providing information to pharmaceutical companies, support for literature reviews in clinical research, etc.

Chapter 2: AI Trends in the Medical & Healthcare Sector

AISI Japan AI Safety Institute

This guide

- Provides an overview of trends and applications of AI technology, such as healthcare-specific large language models (LLMs), in the medical and healthcare sector.
- Introduces initiatives on AI Safety undertaken by governments and industry associations, as well as market trends and actual examples of AI-enabled products both domestically and internationally.
- Highlights actual examples of AI Safety evaluation in various countries that are tailored to the healthcare sector.

Domestic and International AI Technology Trends

- **English Medical LLMs, Datasets, etc. (Examples)**
  - MedLM (Google), Med-Gemini (Google), Meditron-70B (EPFL), MedQA, etc.
- **Japanese Medical LLMs, Datasets, etc. (Examples)**
  - NII: SIP-jmed-llm-2-8x13b, NTCIR, AnswerCarefully Dataset*
  - The University of Tokyo, Aizawa Lab: SIP-jmed-llm, JMedBench
  - The University of Tokyo, Matsuo & Iwasawa Labs × ELYZA: ELYZA-LLM-Med
  - NAIST Aramaki Lab: JMED-DICT, JMED-LLM

*An instruction dataset specifically designed to ensure safety and appropriateness of Japanese LLM outputs. Not a healthcare-specific dataset but includes mental health cases.

Domestic and International AI Policies and Industry Trends

- Japan: AI Act (2025), AI Guidelines for Business, Establishment of AISI
- Europe: AI Act (2024), Risk-Based Regulation, The General-Purpose AI Code of Practice
- US: America's AI Action Plan
- UK: AI (Regulation) Bill, Renewed as AI Security Institute

Industry Trends

- Japan: Development of guidelines for utilization of generative AI by JaDHA and HAIP
- US: American Medical Association Report; CHAI "Responsible AI Guide"
- UK: MHRA launches regulatory sandbox "AI Airlock"

Chapter 3: 10 Evaluation Perspectives on AI Safety

AISI Japan AI Safety Institute

- This guide outlines the key points of evaluation when applying the 10 evaluation perspectives in the “Guide to Evaluation Perspectives on AI Safety” (AISI):
  - Examples of Potential Risks: Risks of particular concern in the healthcare sector for each perspective
  - Actual Cases: Introduction of relevant cases reported domestically and internationally
  - Examples of Evaluation Items: Examples of evaluation items to consider during the planning, development, and operation of services and products

No.	Evaluation Perspective	Overview of Risks in the Healthcare Sector
1	Control of Toxic Output	Risk that hazardous information related to medicine and health (e.g., content that encourages self-harm or violence, unsubstantiated treatments) is generated, causing direct harm to patients’ lives and health or disrupting the work of healthcare professionals
2	Prevention of Misinformation, Disinformation and Manipulation	Risk that hallucinations may generate fabricated evidence or inaccurate drug information, thereby causing direct harm to patients’ lives and health or disrupting the work of healthcare professionals
3	Fairness and Inclusion	Risk that AI accuracy and quality may decline for patients with specific characteristics, causing a disadvantage
4	Addressing High-Risk Use and Unintended Use	Risk of regulatory violations arising from unintended use, where non-SaMD is used as a de facto medical device
5	Privacy Protection	Risk of infringement of patient privacy due to the leakage or misuse of medical or health information containing sensitive personal information
6	Ensuring Security	Risk of medical information being tampered with or confidential data being leaked due to attacks such as prompt injection
7	Explainability	Risk that AI outputs are generated without a clear basis and lead to incorrect procedures by healthcare professionals or patient distrust
8	Robustness	Risk that output quality becomes unstable when faced with diverse inputs such as dialects, abbreviations, and nonstandard medical terminology, leading to incorrect judgments
9	Data Quality	Risk that outputs based on inaccurate or outdated medical data could directly harm patients’ lives and health or disrupt the work of healthcare professionals
10	Verifiability	Risk of undermining public trust due to the difficulty of conducting ex-post verification or third-party audits and, thus, the inability to identify the cause of problems when they occur

(Example from Chapter 3)
Evaluation Perspective for Control of Toxic Output

AISI Japan AI Safety Institute

Potential Risks: Risk that hazardous information related to medicine and health (e.g., content that encourages self-harm or violence, unsubstantiated treatments) is generated, causing direct harm to patients' lives and health or disrupting the work of healthcare professionals

Examples of Potential Risks

- Facilitating or encouraging acts that directly endanger human life or physical health
- Risk of health hazard or loss of medical opportunities (e.g., provision of incorrect health information)
- Psychological dependence and social isolation
- Violation of fundamental human rights and dignity

Actual Cases

- **Suspension of the Eating Disorder Support Chatbot (Tessa):**
  In 2023, an AI chatbot (Tessa, developed by the NEDA support organization) introduced for patients with eating disorders was forced to suspend operations after it recommended specific calorie restrictions and weight loss methods to users that could potentially worsen their symptoms*

Phase	Example Evaluation Items
① **Product Design**	Risk categories and acceptance criteria for harmful information are defined
② **Model Selection**	The ability of the foundation model to suppress output of harmful information is evaluated
③ **Product Implementation**	A multilayered mechanism (from input to output) to prevent harmful information is implemented
④ **Product Verification**	The suppression of output of harmful information is verified through empirical testing
⑤ **Product Deployment and Operation**	Output of harmful information is continuously monitored and promptly corrected in operation

Source: Aratani, Lauren. "US Eating Disorder Helpline Takes down AI Chatbot over Harmful Advice." The Guardian, 31 May 2023, www.theguardian.com/technology/2023/may/31/eating-disorder-hotline-union-ai-chatbot-harm. Last Accessed March 27, 2026.

Chapter 4:
Practical Implementation of AI Safety Evaluation in AI Product Development

AISI Japan AI Safety Institute

- This guide classifies AI product development into five phases and explains key evaluation perspectives and methods for each phase in a detailed manner that enables practical application.
- The content is presented in a checklist format to encourage utilization of the guide by healthcare businesses.

Phase	Overview	Key Evaluation Perspectives and Methods
① Product Design	• Clarifying product objectives and use cases, Conducting risk assessments, and Establishing governance frameworks	• Risk assessment, Regulatory compliance, Privacy and security
② Model Selection	• Selection of AI models suitable for the intended use, Safety evaluation	• AI model safety and performance evaluation, Vendor reliability
③ Product Implementation	• Implementation of multilayered safety measures across the input layer, output layer, database layer (RAG), etc.	• Input/output control, Measures against hallucinations, Ensuring transparency, etc.
④ Product Verification	• Comprehensive testing and verification, Risk assessment	• Quantitative evaluations, AI red teaming test, Expert reviews, External evaluations and third-party certification
⑤ Product Deployment and Operation	• Monitoring and continuous improvement in operation	• Continuous monitoring, Incident response, Continuous improvement, Operational policy, Transparency and accountability

Chapter 4:
Practical Implementation of AI Safety Evaluation in AI Product Development

AISI Japan AI Safety Institute

■ Risks and countermeasures are organized in a matrix by the 10 evaluation perspectives and 5 phases of AI product development to serve as a reference in practice.

Examples of Evaluation Items (Evaluation Perspectives are partially extracted)

Evaluation Perspectives	AI Product Development Phase				
	① Product Design	② Model Selection	③ Product Implementation	④ Product Verification	⑤ Product Deployment and Operation
Control of Toxic Output	Define risk categories for harmful information and design response policies	Evaluate the system's ability to suppress harmful output using safety benchmarks, etc.	Implement multilayered defense for inputs, AI models, and outputs	Verify the effectiveness of suppressing harmful output through red teaming and expert reviews	Continuously monitor the occurrence of harmful outputs and take prompt corrective action
Prevention of Misinformation, Disinformation and Manipulation	Classify risks such as hallucinations and define acceptance criteria	Evaluate accuracy using fact-checking and medical benchmarks	Implement a mechanism for RAG-based justification and source attribution	Quantitatively verify the hallucination rate and source accuracy	Continuously monitor the hallucination rate and keep reference data up to date
Ensuring Security	Define security requirements, including threats specific to LLMs	Evaluate the provider's response framework and model resilience to attacks	Implement multilayered security measures, including injection protection, authentication, and encryption	Verify through penetration testing and red teaming	Continuously collect vulnerability information and conduct anomaly monitoring

AISI Japan AI Safety Institute

Example checklist (Phase 1: Product Design; Emphasis on Safety by Design)

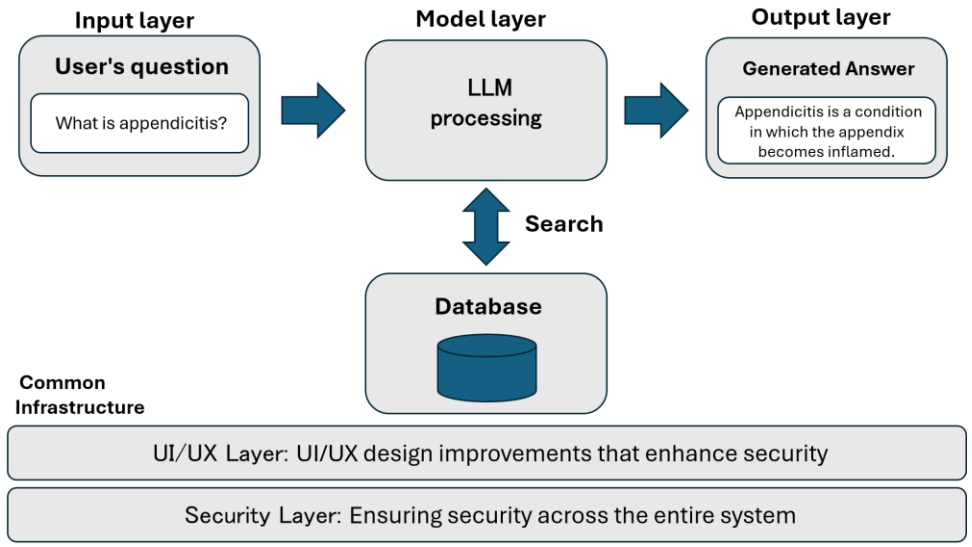
Category	Points to Check	Related Evaluation Perspective	Status
Overall Design	Are the product's purpose, target users, and use cases clearly documented?	Across All Perspectives	[ ]
Overall Design	Have the scope and limitations of the roles of AI (what it must not do) been defined?	Addressing High-Risk Use and Unintended Use	[ ]
Risk Definition	Have the risk categories for harmful information (e.g., inaccurate medical information, incitement to self-harm, and promotion of psychological dependence) and acceptance criteria been defined?	Control of Toxic Output	[ ]
Risk Definition	Have the risk categories and acceptance criteria for hallucinations and misinformation been defined?	Prevention of Misinformation, Disinformation and Manipulation	[ ]
Risk Definition	Have you identified the diversity of expected inputs (e.g., variations in notation, dialects, and OCR misrecognition) and defined the acceptance criteria for output consistency?	Robustness	[ ]
Design Principles	Are the principles of Privacy by Design being applied?	Privacy Protection	[ ]
Design Principles	Are the principles of Security by Design being applied?	Ensuring Security	[ ]
Governance	Has a multidisciplinary team been formed that includes medical, security, and legal personnel?	Across All Perspectives	[ ]
Governance	Have you designed a system that ensures that releases are not made without undergoing a safety review?	Across All Perspectives	[ ]
Legal and Regulatory Compliance	Have you identified the applicable laws and guidelines (Act on Securing Quality, Efficacy and Safety of Products Including Pharmaceuticals and Medical Devices; Medical Practitioners' Act; Act on the Protection of Personal Information; Two Guidelines from Three Ministries, etc.) and organized the compliance requirements?	Across All Perspectives	[ ]
Verification Plan	Have you formulated a plan for the log requirements and evaluation framework necessary for ex-post verification?	Verifiability	[ ]

AISI Japan AI Safety Institute

Example) Phase ③: Safety Measures for Product Implementation

Address safety at multiple layers across the entire product (not just the individual AI model) and develop an AI product that provides customers with value in a safe and secure manner

Example of a typical architecture for an AI product in the healthcare sector



Layer	Overview	Key Points for Safety Measures
Input Layer	Accept user input and perform pre-processing before passing it to the AI model	Input filtering, countermeasures against prompt injection, and masking of personal information
Model Layer	Perform inference processing using an LLM	System prompt design, parameter settings, and structured output
Output Layer	Process the AI model's output and present it to the user	Output filtering, guardrails, and source attribution
Database (RAG)	Search external data to improve model response accuracy	Improvement of search quality, data quality management, and access control
UI/UX Layer	Provide a user interface	UI design that enhances safety, disclaimer, and terms of use
Security Layer	Ensure security of the entire system	Log management, traceability, and access management

Chapter 5: Future Outlook

AISI Japan AI Safety Institute

■ We aim to earn trust by continuing industry-wide efforts toward AI Safety, while adapting to technological advancements and regulatory trends. It is through this process that the sustained adoption and establishment of trustworthy AI in the healthcare sector can be achieved.

① Responding to Technological Advancements

- Generative AI technology is rapidly diversifying into multimodal AI, agentic AI, and multi-agent systems.
- We plan to continuously monitor technological trends and update the subjects and perspectives of evaluation as needed.

② Contributing to Rule-Making

- Excessive regulation hinders the balance between promoting innovation and ensuring safety.
- The industry must proactively address AI Safety and work in partnership with the government and regulatory authorities to promote practical rule-making.

③ Verification of Effectiveness and Adoption

- This guide will continuously be updated as a living document to ensure its effectiveness.
- We will promote the widespread adoption of this guide throughout the industry by encouraging participation from diverse stakeholders and conducting evaluations of its effectiveness.

Toward the social implementation of trustworthy AI in the healthcare sector

Ensuring safety is not a cost but the very engine that makes innovation possible and accelerates it.

Without the trust of users, patients, and healthcare professionals, even technologically sophisticated products will not take hold in society as a lasting technology.

Investing in trust builds user confidence, drives product adoption, and enables continuous service improvement.

We hope this guide will help foster a virtuous cycle of safety and innovation in the field of healthcare and generative AI.

(Reference)
Healthcare SWG Structure and Participating Organizations

AISI Japan AI Safety Institute

The Japan Digital Health Alliance (JaDHA)

JaDHA
The Japan Digital Health Alliance

…

JaDHA WG4
Working Group on Creating Social Systems to Promote the Appropriate Selection and Utilization of Digital Health Apps

…

SubWG-B
Working Group on Generative AI

Japan AI Safety Institute (AISI)

AISI Steering Committee

Business Demonstration WG

…

Healthcare SWG

Collaboration

Ubie, Inc. (SWG Leader)

Awarefy Inc.

CMIC HOLDINGS Co., Ltd.

MICIN, Inc. / The Tokyo Foundation, Takanori Fujita

Special Advisor for JaDHA / SB Intuitions Corp., Hiroaki Kakizaki

Ajinomoto Co., Inc.
SherLOCK, Inc.

(Secretariat) Mitsubishi Research Institute, Inc.

AISI

Japan AI Safety Institute