EMA Lab
Arisa EMA | The University of Tokyo

Transparency and Integrity in HAIP Reporting Framework
- Insights and Recommendations for AISI Network -

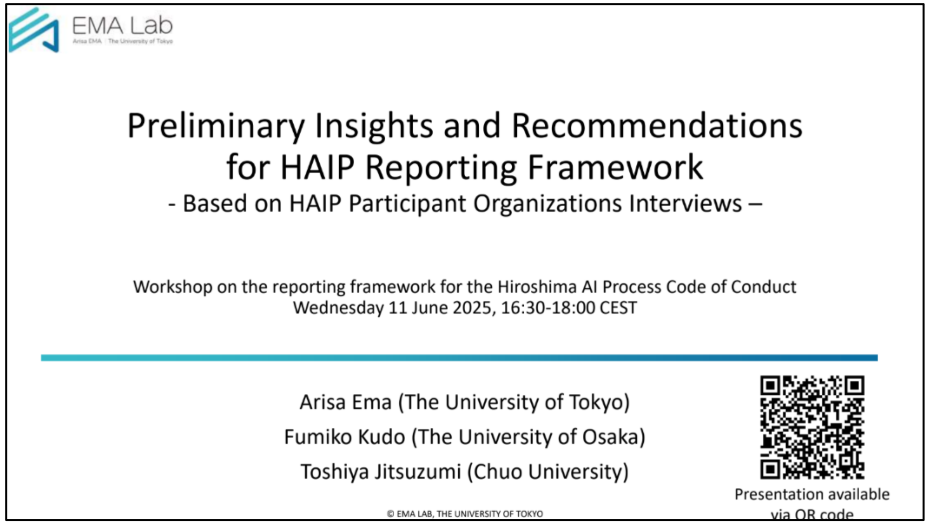
Future Developments of HAIP: Initial Reporting Outcomes and Alignment with the Japan AI Act:
Side event to the AISI International Network Meeting, Vancouver, Canada

18 July 2025

Arisa Ema (The University of Tokyo / Japan AISI)
Fumiko Kudo (The University of Osaka)
Toshiya Jitsuzumi (Chuo University)

Presentation available via QR code

Interviews with HAIP Participant Companies



EMA Lab

Preliminary Insights and Recommendations for HAIP Reporting Framework

- Based on HAIP Participant Organizations Interviews –

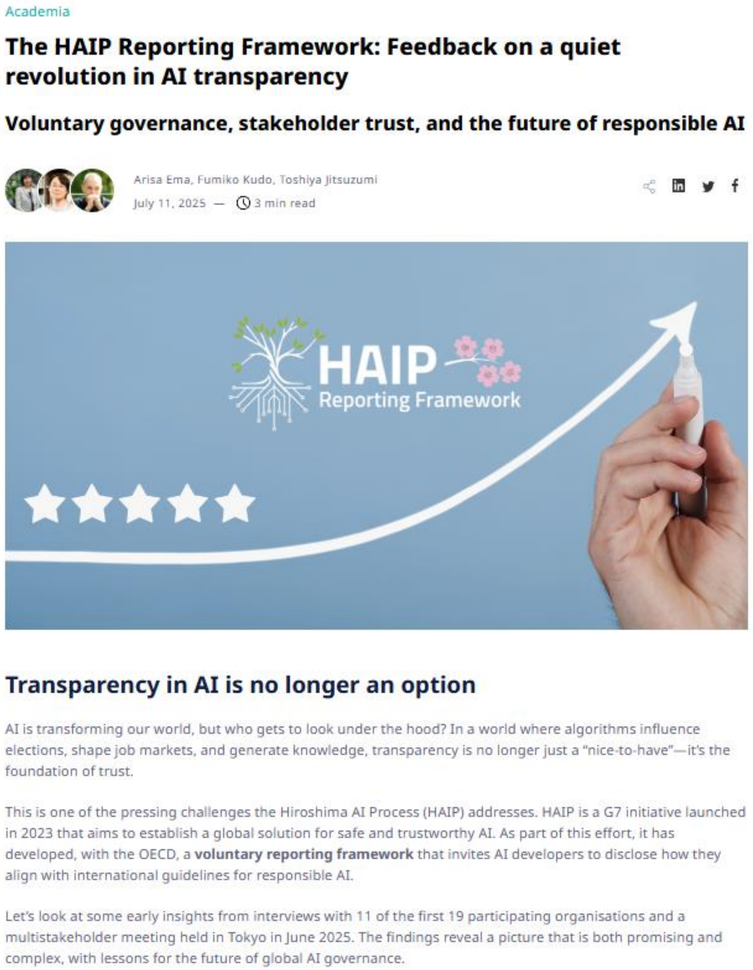
Workshop on the reporting framework for the Hiroshima AI Process Code of Conduct
Wednesday 11 June 2025, 16:30-18:00 CEST

Arisa Ema (The University of Tokyo)
Fumiko Kudo (The University of Osaka)
Toshiya Jitsuzumi (Chuo University)

Presentation available via QR code

© EMA LAB, THE UNIVERSITY OF TOKYO

https://drive.google.com/file/d/1Re8fApWZTVzU1xMXBBuS3V-C3pASU7Kz/view



Academia

The HAIP Reporting Framework: Feedback on a quiet revolution in AI transparency

Voluntary governance, stakeholder trust, and the future of responsible AI

Arisa Ema, Fumiko Kudo, Toshiya Jitsuzumi
July 11, 2025 — 3 min read

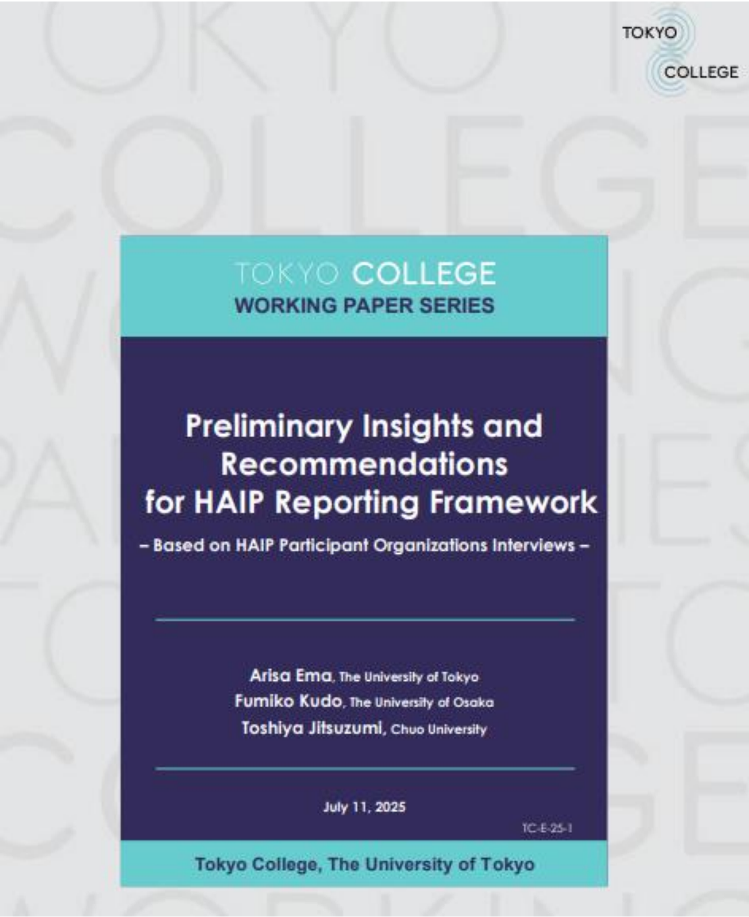
Transparency in AI is no longer an option

AI is transforming our world, but who gets to look under the hood? In a world where algorithms influence elections, shape job markets, and generate knowledge, transparency is no longer just a “nice-to-have”—it’s the foundation of trust.

This is one of the pressing challenges the Hiroshima AI Process (HAIP) addresses. HAIP is a G7 initiative launched in 2023 that aims to establish a global solution for safe and trustworthy AI. As part of this effort, it has developed, with the OECD, a **voluntary reporting framework** that invites AI developers to disclose how they align with international guidelines for responsible AI.

Let’s look at some early insights from interviews with 11 of the first 19 participating organisations and a multistakeholder meeting held in Tokyo in June 2025. The findings reveal a picture that is both promising and complex, with lessons for the future of global AI governance.

https://oecd.ai/en/wonk/the-haip-reporting-framework-feedback-on-a-quiet-revolution-in-ai-transparency



TOKYO COLLEGE

TOKYO COLLEGE
WORKING PAPER SERIES

Preliminary Insights and Recommendations for HAIP Reporting Framework

– Based on HAIP Participant Organizations Interviews –

Arisa Ema, The University of Tokyo
Fumiko Kudo, The University of Osaka
Toshiya Jitsuzumi, Chuo University

July 11, 2025

TC-E-25-1

Tokyo College, The University of Tokyo

https://www.tc.u-tokyo.ac.jp/blog/wp-content/themes/tokyocollege/publication/WP_TC-E-25-1_Ema.pdf

Who Are the Target Audiences for HAIP Reporting?

Audience Type	Description	Typical Motivation
International Bodies	G7 / OECD Partners	- Visibility in AI governance - International alignment
Policy Stakeholders	Government bodies, regulators	- Gain trust - Influence on regulatory frameworks
Business & Technical Partners	B2B clients, external developers, corporate partners	- Contractual clarity - Risk accountability
General Public	Shareholders, citizens, job-seeking students	- Trust-building - Brand strategy
Internal teams	Employees	- Create internal alignment and awareness on AI governance

What Effort Did HAIP Participation Require from Organizations?

- Reorganization of existing info vs. creation of new materials
  - Internal practices were sometimes not documented, structured for external audiences

- Internal approval hurdles (especially for Japanese companies)
  - Convincing internal teams of why transparency reporting matters
  - The submission deadline coinciding with fiscal year-end in March (in Japan)
  - Desire for broader understanding of HAIP’s purpose and brand

Ambiguities & Misunderstandings in the HAIP Questionnaires

- Ambiguities in:
  - Scope: Is the question referring to a specific AI system or company policies?
  - Role: Should we answer as a developer, a provider or both?
    - In B2Bcases, disclosure to clients can be particularly sensitive or difficult.
  - Audience: Is the report for government, clients, or the public?

- Needed for clearer templates or examples
  - However, there are tension between flexibility and clarity

Report should Promote Transparency – Not Scoring

- Many companies assert that HAIP reporting framework should not be used for ranking/scoring without considering business model differences
  - Submitting a report demonstrates a commitment to transparency and responsible AI — this act itself should be encouraged
  - Integrity matters – need to prohibit unfair or deceptive acts or practices

Integrity matters – pre and post

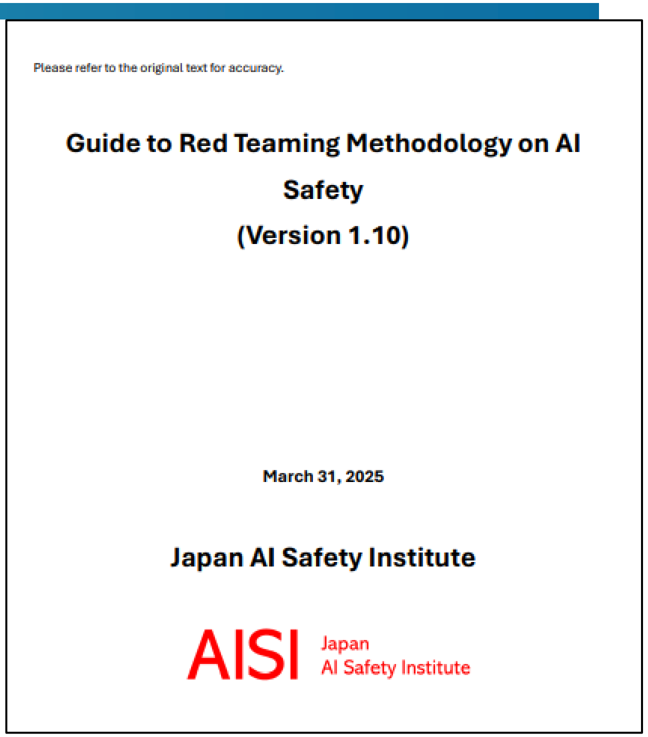
- Entities responsible for ensuring the integrity of the HAIP report

Level	Function	Actor
Expert Guidance, Advice and Support (pre and post)	Help companies write accurate report	AISI network / OECD-GPAI / UN
Oversight and Monitoring	Prohibit unfair or deceptive acts or practices	Government agencies / Authorities / Courts
Social Accountability	Detect and deter false claims and raise literacy	Civil society / Market / Journalism / Academia

Expected Roles for AISI in Supporting HAIP

- Technical Expert advisory
  - Identify and disseminate good practice
  - Provide pre- and post-report advisory sessions
  - Offer templates and tutorials
  - Serve as trusted consultation point

- Japan AISI already published Red Teaming Guidance
  - It could extend to report writing support!



Please refer to the original text for accuracy.

Guide to Red Teaming Methodology on AI Safety
(Version 1.10)

March 31, 2025

Japan AI Safety Institute

AISI Japan AI Safety Institute

https://aisi.go.jp/output/output_framework/guide_to_red_teaming_methodology_on_ai_safety/

- AISI network could help disseminate good practices and guide companies in choosing and explaining their approaches with integrity
  - This will be a support for SMEs as well

HAIP Section	Key Technical Aspects required
1 Risk Identification and Evaluation	• Conducting **technical testing** (e.g. red-teaming, penetration test) to assess AI system readiness before deployment • Identifying vulnerabilities, misuse through **adversarial testing**
2 Risk Management and Information Security	• Performing **testing** in secure, isolated or sandboxed environments • Implementing robust **cybersecurity risk assessments** • Protecting proprietary AI elements (e.g., model weights, algorithms) through **access controls and encryption**
3 Transparency Reporting on Advanced AI Systems	• Publicly disclosing **detailed results of technical evaluations** • Providing information on **model capabilities, limitations, and appropriate use domains** derived from technical assessments
4 Organizational Governance, Incident Management and Transparency	• -
5 Content Authentication & Provenance Mechanisms	• Developing and implementing **technical mechanisms** (e.g., watermarking, metadata tagging, digital signatures) to identify AI-generated content • Adhering to **international technical standards and best practices** for content provenance
6 Research & Investment for AI Safety and Risk Mitigation	• Investing in and conducting **research to develop new technical evaluation methods and tools** for AI safety, security, and trustworthiness. • Advancing research in areas like **bias detection, disinformation, robustness, and explainability** through technical means
7 Advancing Human and Global Interests	• -

Next steps

- Shared Goals
  - Promote transparency in AI governance
  - Improve comparability across reports
  - Preserve flexibility and adaptability for diverse actors

- Next stems
  - Our detailed report and recommendations will be compiled this summer
  - We welcome feedback and continued dialogue from all stakeholders

Special thanks

We sincerely thank the following organizations and individuals for their cooperation in the interview process:

Organizations (by submission order):

KDDI Corporation, SoftBank Corp., Preferred Networks, NEC Corporation, NTT, Microsoft, Salesforce, Anthropic, OpenAI, Google, Fujitsu, Rakuten Group

Additional organizations were invited, and we look forward to including their input in future versions.

Presentation available via QR code