Discussion Paper

AMAIS: Activity Map on AI Safety

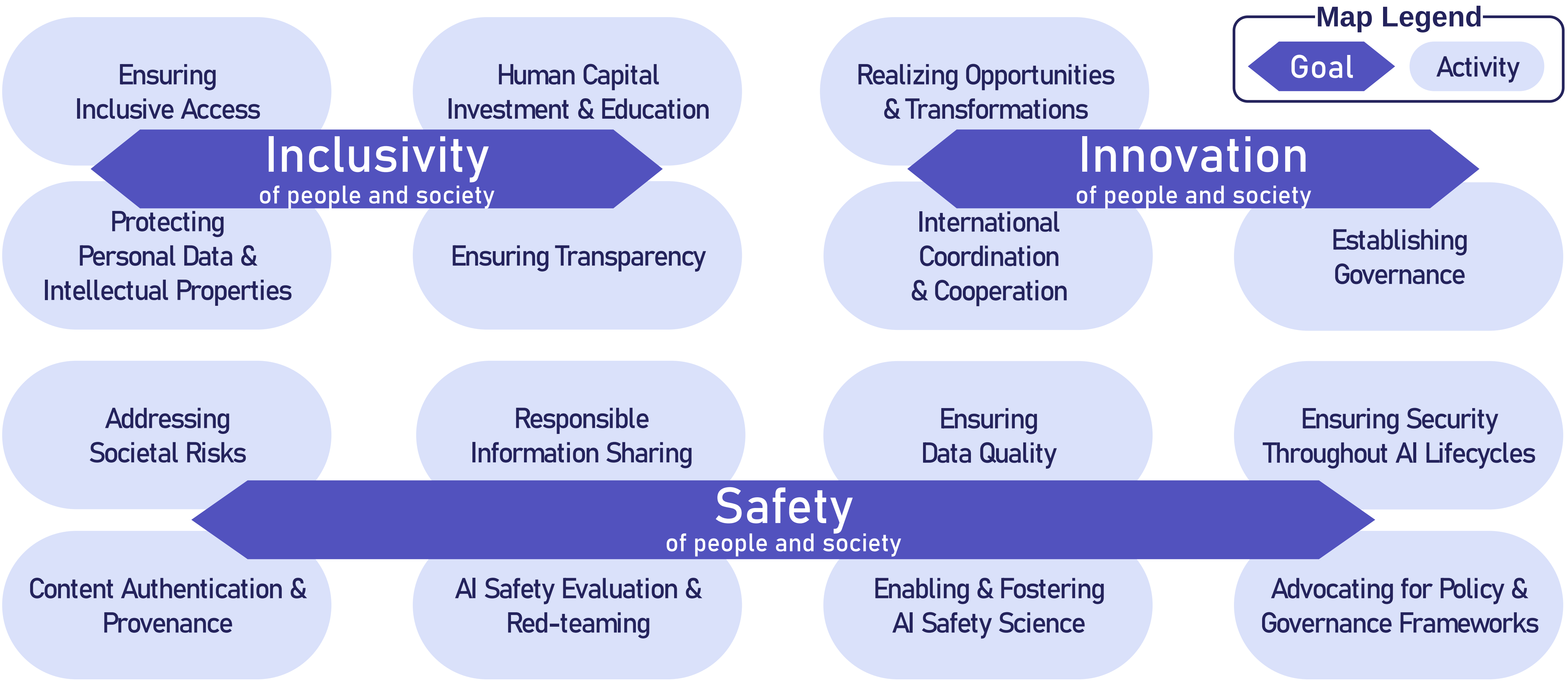
Feb. 2025

AISI Japan
AI Safety Institute

Activity Map on AI Safety (AMAIS)

AISI Japan AI Safety Institute

AMAIS shown below provides a comprehensive overview of AI safety activities, supporting discussion on their scope and priorities.



Map Legend

Goal

Activity

Ensuring Inclusive Access

Human Capital Investment & Education

Inclusivity of people and society

Protecting Personal Data & Intellectual Properties

Ensuring Transparency

Realizing Opportunities & Transformations

Innovation of people and society

International Coordination & Cooperation

Establishing Governance

Addressing Societal Risks

Responsible Information Sharing

Ensuring Data Quality

Ensuring Security Throughout AI Lifecycles

Safety of people and society

Content Authentication & Provenance

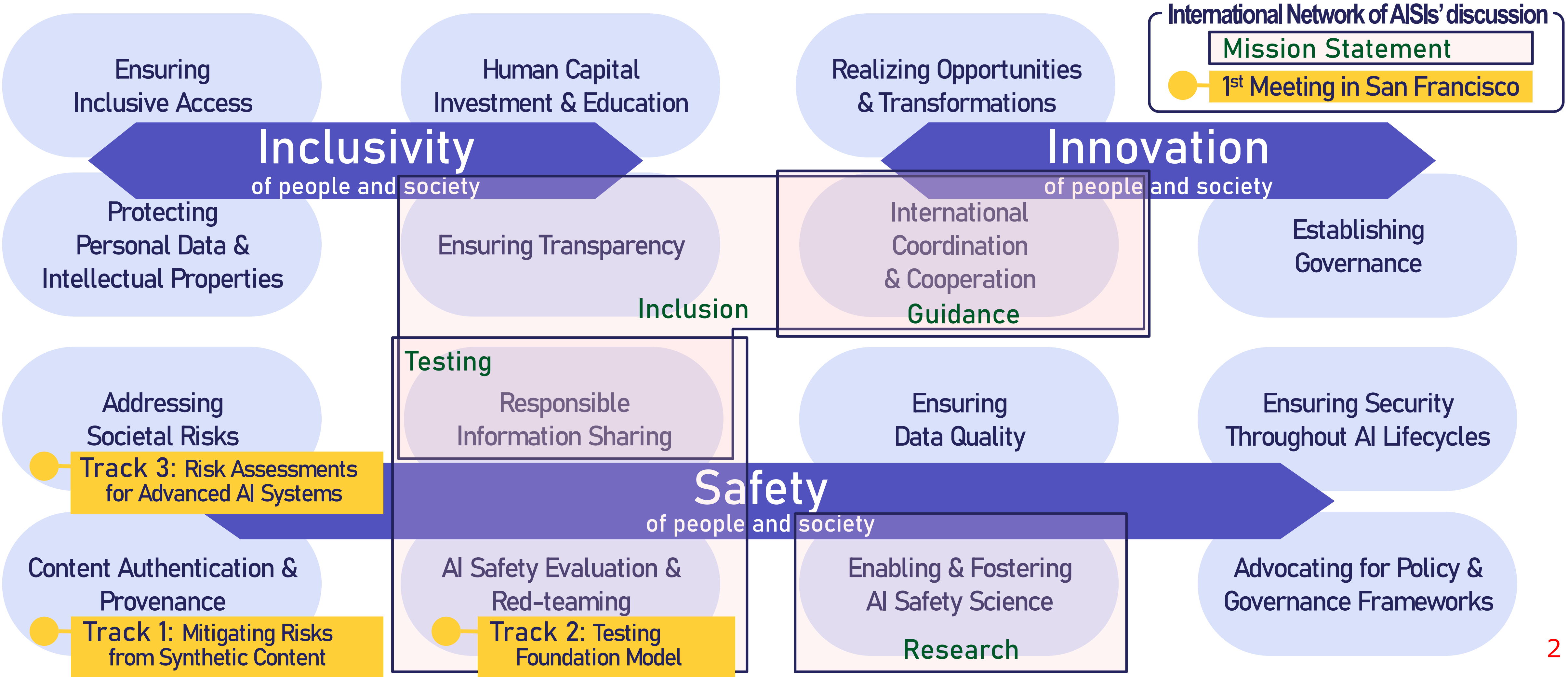
AI Safety Evaluation & Red-teaming

Enabling & Fostering AI Safety Science

Advocating for Policy & Governance Frameworks

Ongoing Actions on AMAIS

AMAIS captures the actions being discussed by the international AI communities, including the International Network of AI Safety Institutes, as shown below.



International Network of AISIs' discussion

Mission Statement

● 1st Meeting in San Francisco

Ensuring Inclusive Access

Human Capital Investment & Education

Realizing Opportunities & Transformations

Inclusivity of people and society

Innovation of people and society

Protecting Personal Data & Intellectual Properties

Ensuring Transparency

Inclusion

International Coordination & Cooperation

Guidance

Establishing Governance

Testing

Addressing Societal Risks

● Track 3: Risk Assessments for Advanced AI Systems

Responsible Information Sharing

Ensuring Data Quality

Ensuring Security Throughout AI Lifecycles

Safety of people and society

Content Authentication & Provenance

● Track 1: Mitigating Risks from Synthetic Content

AI Safety Evaluation & Red-teaming

● Track 2: Testing Foundation Model

Enabling & Fostering AI Safety Science

Research

Advocating for Policy & Governance Frameworks

Summary

✓ This paper shows the “**Activity Map on AI Safety (AMAIS)**,” which is a whole picture of the activities required to achieve AI safety based on the outcomes of the “Hiroshima AI Process” and the “AI Seoul Summit.”

✓ **The terminology challenges**, which were identified during the process of creating AMAIS, are also shown in this paper.

✓ As the next step, we will hold a technical discussion on the need for both **the activity map to explore the scope and priorities of AI safety**, and **the terminology as a common language**, with the aim of **promoting AI safety for innovation and inclusivity** in a more robust and formalized fashion.

This paper is a living document to promote international discussions on AI safety. This paper does **not** indicate any activity that should be undertaken by AISIs or other organizations in each country.

Contact: aisi-amais-info [at] ipa [dot] go [dot] jp

Sources of Activities

AMAIS was created based on the following items.

	Activity	Sources
Safety	Content Authentication & Provenance	Hiroshima AI Process[1] - Guiding Principles[3] 1, 7
	Addressing Societal Risks	Hiroshima AI Process[1] - Guiding Principles[3] 8, Code of Conduct[4] 1
	AI Safety Evaluation & Red-teaming	Hiroshima AI Process[1] - Guiding Principles[3] 1
	Responsible Information Sharing	Hiroshima AI Process[1] - Guiding Principles[3] 2, 4
	Enabling & Fostering AI Safety Science	Hiroshima AI Process[1] - Guiding Principles[3] 8, AI Seoul Summit 2024[5] - Seoul Declaration[6] 1, 4, Seoul Statement[7] 2
	Ensuring Security Throughout AI Lifecycles	Hiroshima AI Process[1] - Guiding Principles[3] 6
	Advocating for Policy & Governance Frameworks	AI Seoul Summit 2024[5] - Seoul Declaration[6] 6
	Ensuring Data Quality	Hiroshima AI Process[1] - Code of Conduct[4] 11
Inclusivity	Protecting Personal Data & Intellectual Property	Hiroshima AI Process[1] - Guiding Principles[3] 11
	Ensuring Inclusive Access	AI Seoul Summit 2024[5] - Seoul Declaration[6] 5, 6
	Ensuring Transparency	Hiroshima AI Process[1] - Guiding Principles[3] 3, Code of Conduct[4] 3
	Human Capital Investment & Education	Hiroshima AI Process[1] - Guiding Principles[3] 9
Innovation	International Coordination & Cooperation	Hiroshima AI Process[1] - Guiding Principles[3] 10, AI Seoul Summit 2024[5] - Seoul Declaration[6] 5, 7, Seoul Statement[7] 2,3
	Realizing Opportunities & Transformations	Hiroshima AI Process[1] - G7 Leaders' Statement[2]
	Establishing Governance	Hiroshima AI Process[1] - Guiding Principles[3] 5, AI Seoul Summit 2024[5] - Seoul Declaration[6] 3, 6

Activities and Relevant Terms

AISI Japan AI Safety Institute

Examples of terms that often appear in discussions of each activity.

	Activity	Relevant Terms
Safety	Content Authentication & Provenance	Originator Profile, Disinformation, Hallucination, Watermarking, Synthetic Contents, Provenance mechanisms, Disclaimer, AI Label
	Addressing Societal Risks	Dual-use, Foundation Model, Artificial General Intelligence (AGI), AI-agent, General Purpose AI (GPAI), Risk management for CBRN, AI for Critical Infrastructure, IT/OT, Cognitive and Behavioral Manipulation, Profiling, Job Market in the age of AI
	AI Safety Evaluation & Red-teaming	Threat Actor Uplift Evaluation, External Testing, Automated Evaluation, Test-bed, Robustness, Alignment
	Responsible Information Sharing	Bounty Program, Multi-Stakeholder, ‘Incident Response and Sharing among Industry, Academia and Government,’ Early Warning Information Sharing, Incident Report
	Enabling & Fostering AI Safety Science	Academic Research, Grants and Startups by Government, Safety for Emerging Technology, Foundation Model
	Ensuring Security Throughout AI Lifecycles	Cyber, Physical Access Control, Information Security, Risk Mitigation, Internal Threat Detection Program, Security for AI, AI for Security
	Advocating for Policy & Governance Frameworks	Developing Guidelines, Identifying Value-chain, Addressing AI Safety Washing, Ensuring Fair Competition, Certification System, Taxonomy and Terminology
	Ensuring Data Quality	Traceability, Output Attribution, Enhancing Interpretability
Inclusivity	Protecting Personal Data & Intellectual Property	Privacy, Copyright, Safeguard
	Ensuring Inclusive Access	Accessibility, Safety Net, Diversity, Outreach, Human Welfare, Protection from Disasters
	Ensuring Transparency	Responsible AI Development, Ethics, Trustworthiness, Accountability, Fairness, Transparency Report, Model Card, System Card, Data Card, Human-Centric
	Human Capital Investment & Education	Outreach, Certification System, School Education
Innovation	International Coordination & Cooperation	Interoperability, Guardrail, Standards Development Organizations, Cross-border, Joint Testing, Cross-disciplinary, Scientific
	Realizing Opportunities & Transformations	Public Sector, Manufacturing, Robotics and Mobility Logistics and Healthcare, Government, SMEs and Startups
	Establishing Governance	Risk Management, Management System, Risk Assessment, Accountability

Terminology Challenges

While there are many AI terminologies and taxonomies[7-17], we found several challenges during the creation of AMAIS. **Ideally, each term would have exactly one clear definition**, but that is **not the current reality**. Here, we illustrate these challenges with examples, thereby raising topics for future discussion. They are not necessarily separate but are **often intricately intertwined**.

1. **Vague Definitions (e.g., AI, AI Agent)**
   Some terms are left undefined or only vaguely defined, which can lead to confusion in discussions.
2. **Overlapping Meanings for a Single Term (e.g., Red Teaming in Cybersecurity/AI Safety)**
   A single term can have overlapping meanings chosen depending on the context. Therefore, it's crucial to understand such terms in their specific context.
3. **One Meaning Expressed by Multiple Terms (e.g., Supervised Fine-tuning/Instruction Tuning)**
   Some terms may essentially mean the same thing. Using them at the same time can lead to misunderstandings about the discussion or imply a distinction that doesn't actually exist.
4. **Similar Yet Distinct Terms (e.g., Misinformation/Disinformation/Malinformation, GPAI/AGI)**
   Some terms looks like similar but have different meanings, and if you don't know the difference, you could easily get confused.

References (International Declarations)

1. Hiroshima AI Process, https://www.soumu.go.jp/hiroshimaaiprocess/en/index.html
2. G7 Leaders' Statement on the Hiroshima AI Process, https://www.soumu.go.jp/hiroshimaaiprocess/pdf/document01_en.pdf
3. Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems, https://www.soumu.go.jp/hiroshimaaiprocess/pdf/document04_en.pdf
4. Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems, https://www.soumu.go.jp/hiroshimaaiprocess/pdf/document05_en.pdf
5. AI Seoul Summit 2024, https://www.gov.uk/government/topical-events/ai-seoul-summit-2024
6. Seoul Declaration for safe, innovative and inclusive AI by participants attending the Leaders' Session, https://www.gov.uk/government/publications/seoul-declaration-for-safe-innovative-and-inclusive-ai-ai-seoul-summit-2024/seoul-declaration-for-safe-innovative-and-inclusive-ai-by-participants-attending-the-leaders-session-ai-seoul-summit-21-may-2024
7. Seoul Statement of Intent toward International Cooperation on AI Safety Science, https://www.gov.uk/government/publications/seoul-declaration-for-safe-innovative-and-inclusive-ai-ai-seoul-summit-2024/seoul-statement-of-intent-toward-international-cooperation-on-ai-safety-science-ai-seoul-summit-2024-annex

References (Terminologies & Taxonomies)

7. EU-U.S. Terminology and Taxonomy for Artificial Intelligence, https://digital-strategy.ec.europa.eu/en/library/eu-us-terminology-and-taxonomy-artificial-intelligence
8. Creation of a taxonomy for the European AI ecosystem, https://eit.europa.eu/library/creation-taxonomy-european-ai-ecosystem
9. AI Use Taxonomy: A Human-Centered Approach, https://www.nist.gov/publications/ai-use-taxonomy-human-centered-approach
10. AI Taxonomy, https://www.southampton.ac.uk/publicpolicy/support-for-researchers/policy%20briefs/policy%20briefs/ai-taxonomy.page
11. A data for AI taxonomy, https://theodi.org/news-and-events/blog/a-data-for-ai-taxonomy/
12. AI Risks Taxonomy, https://unidir.org/wp-content/uploads/2023/10/UNIDIR_Research_Brief_AI_International_Security_Understanding_Risks_Paving_the_Path_for_Confidence_Building_Measures.pdf
13. A Taxonomy of Trustworthiness for Artificial Intelligence, https://cltc.berkeley.edu/wp-content/uploads/2023/01/Taxonomy_of_AI_Trustworthiness.pdf
14. AI Mapβ 2.0, https://www.ai-gakkai.or.jp/pdf/aimap/AIMap_EN_20210901.pdf
15. Guide to Evaluation Perspectives on AI Safety, https://aisi.go.jp/effort/effort_framework/guide_to_evaluation_perspective_on_ai_safety
16. Guide to Red Teaming Methodology on AI Safety, https://aisi.go.jp/effort/effort_framework/guide_to_red_teaming_methodology_on_ai_safety
17. Crosswalk 1 – Terminology NIST AI Risk Management Framework (NIST AI RMF) and Japan AI Guidelines for Business (AI GfB), https://aisi.go.jp/assets/pdf/AISI_Crosswalk1_RMF_GfB_ver1.0.pdf

AISI

Japan AI Safety Institute