JAPAN AI SAFETY WORKSHOP

July 5, 2025

Summary of the Discussions and Preliminary Insights from the 1st international AI Safety event held in Japan with the backing of Japan AI Safety Institute. The event was organized by Henkaku Center and Digital Garage.

Merve Hickok,
Visiting AI Researcher (Henkaku Center – Chiba Tech University)
President & Policy Director (Center for AI and Digital Policy)

Contents

Introduction

Open dialogue on AI safety is essential for building trust among users, stakeholders, and the broader public. By addressing potential risks transparently and collaboratively, organizations can foster a sense of accountability and demonstrate their commitment to responsible AI development. Such transparency is crucial not only for alleviating fears but also for ensuring widespread adoption, as stakeholders are far more likely to embrace AI systems they believe are designed and governed with care for human values and societal wellbeing. In this context, ongoing conversations about AI safety form the cornerstone of establishing the confidence necessary for AI technologies to flourish responsibly.

Japan occupies a special place in AI policy and governance. Starting almost a decade ago, Japanese researchers and policymakers initiated many international dialogues on the necessity of AI governance and international collaboration. Japan was influential for the start of these conversations in G7 and Organisation for Economic Co-operation and Development (OECD) in 2016, which eventually led to the OECD AI Principles in 2019. Japan has contributed to the formulation of the UNESCO Recommendation on the Ethics of AI (adopted by 193 countries in 2021), as well as the drafting of the Council of Europe's Framework Convention on Artificial Intelligence and Human Rights, Democracy, and the Rule of Law (the first international treaty on AI). When more advanced AI systems were introduced in 2023, it was again Japanese policymakers who lead the formulation and consensus-building on Hiroshima AI Process Comprehensive Policy Framework (which includes a list of guiding principles for all AI actors, voluntary code of conduct for organizations developing advanced AI systems, and a reporting mechanism for both developers and deployers of advanced AI systems). To support the research and project-based coordination with Global Partnership on AI (GPAI) and other organizations, Japan established the GPAI Tokyo Expert Support Center to support research. Immediately after the counterparts in the United States and United Kingdom, Japan also established one of the first AI Safety institutes in February 2024. Japan AI Safety Institute is already an important source of research and knowledge on the subjects most relevant to AI safety concerns. Most recently, in

May 2025, Japan enacted the “Act on the Promotion of Research, Development, and Utilization of Artificial Intelligence-Related Technologies” to encourage the adoption of AI by businesses and society, while ensuring AI systems are safe and trustworthy.

Within this larger context, July 5th marked the first international AI safety event held in Japan with the backing of Japan AI Safety Institute. The event was organized by Henkaku Center and Digital Garage, as part of the Chiba Tech Henkaku Center Symposium. The Symposium, themed “Design and Science,” highlighted innovative anti-disciplinary research and ongoing projects at the Center related to AI, Neurodiversity, and Education in Japan.

The overarching theme of the AI Safety Workshop was AI safety, governance and quality management in Japan. This was the first event in what the organizers envisage as a series of community building and engagement activities focused on these subjects. With the series of AI safety workshops, events, and continuous dialogue in between, the Henkaku Center aims to help build and strengthen a community of AI safety professionals and practitioners and hopefully bridge some of the existing efforts in Japan and beyond.

The community element is important as Henkaku Center believes in the synergy and power of such connections, and the importance of diverse perspectives to drive change. The Center was established in 2021 to “bring together researchers from all sectors of society and both the technical and cultural disciplines to imagine, design, architect, and build technical platforms as well as cultural output to help society through the radical transformation that is happening.” Not all problems can be solved with AI. AI should not be a hammer slammed on every problem; instead, it should be carefully considered as one tool among many for addressing challenges. When AI is applied indiscriminately, it can lead to unintended consequences, overlook context-specific needs, and erode public trust. However, AI can indeed bring many benefits and AI safety is one of the most critical elements to drive positive transformation in the deployment of artificial intelligence across society. Ensuring that AI systems are designed with rigorous safety standards helps mitigate risks, fosters transparency, and promotes responsible innovation. When organizations prioritize AI safety, they build a foundation of trust that enables collaboration among stakeholders, policymakers, and the public. This trust is not only essential for addressing concerns about unintended consequences and ethical dilemmas but also paves the way for the broader adoption of AI technologies. Ultimately, a strong commitment to AI safety supports the sustainable growth of AI and ensures its benefits are realized by individuals, businesses and communities.

AI safety requires involvement of a variety of stakeholders to address the many challenges and concerns. The Workshop participants will hopefully benefit from each other’s perspectives and knowledge. The first AI Safety Workshop included representatives from Japanese corporations, startups, and academia to discuss AI as a critical technology for our

societies and businesses. AI is also a borderless technology, and there is a lot to share and collaborate with international partners. For that reason, Japanese colleagues were joined by international researchers and practitioners from the Republic of Korea, Singapore, United States, United Kingdom, Germany and Netherlands. Finally, the Workshop also brought together partners from the OECD and Japanese government with regards to the latest developments on the Hiroshima AI Process Framework.

In summary, the AI Safety Workshop had a well-balanced representation of participants in terms of large traditional companies vs the startups, business stakeholders vs academia and government; Japanese practitioners vs international ones, and finally the technical professionals vs those working in governance of AI technologies.

Presentations

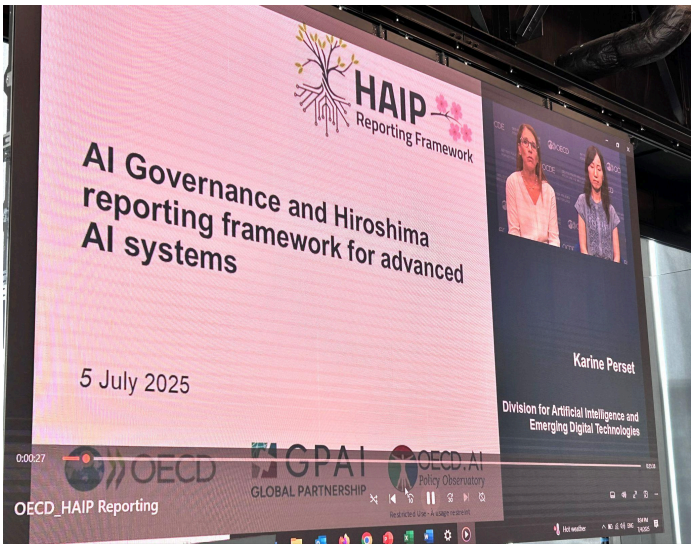
Merve Hickok, as one of the main organizers and facilitators of the AI safety initiative, delivered a short remark to welcome the participants to the Workshop and reiterated the objectives of the initiative.

AI Safety Workshop opened with a welcome keynote from **Akiko Murakami**, the Executive Director of the Japan AI Safety Institute (J-AISI). J-AISI was launched in February 2024, and since then it has become an integral part of the AI safety conversations both domestically and internationally. J-AISI's role as an agency is to support public and private sector initiatives to promote the safe and secure use of AI. It achieves this via *Supporting the government* (investigating AI safety, examining evaluation methods, and creating standard), *Becoming a Hub of AI Safety in Japan* (collecting the latest industry-academia initiatives; promoting collaboration among related entities; collaborating with international AI safety institutions), and *Collaborating with AI Safety-related organizations* (collaborating with national research institutes; promoting partnerships).

In her keynote, Director Murakami asked whether the fear of AI is our greatest enemy. She warned that the biggest risk with AI may not be its dangers, but the loss of opportunities if we do not use it and miss out on important gains in efficiency and driving new innovation. However, she also noted that this does not mean we should disregard safety concerns. In fact, her closing remark was "innovation can only accelerate when safety is ensured." Referring to how the brakes in a car allow the driver to trust

the product and accelerate, she called all to **advance innovation and adoption of AI by integrating safety.** Her message was repeated in every presentation that followed the opening remarks.

The next presentation was a recorded one by **Karine Perset**, the Acting Head of the OECD AI and Emerging Digital Technologies Division. Perset walked the participants through the Hiroshima AI Process Reporting Framework (HAIP: https://transparency.oecd.ai) – its objectives and current state. Companies developing and deploying advanced AI systems voluntarily submit public information on their AI governance activities through a standard template format. As of July 2025, 20 organizations submitted their reports. The HAIP reports provide detailed information on topics such as risk assessment methodologies, risk classification and management, model transparency, accountability mechanisms, and socio-technical impacts. The reports provide comparative insights to drive the implementation of AI safety and governance. Perset's presentation provided a preliminary analysis of the 20 public reports and the common threads across organizational practices. These early trends included management of risk across the lifecycle of AI systems by using established frameworks (NIST RMF, ISO standards etc), and governance methods such as secure testing environments, red teaming, access controls, specialized AI governance teams, ethics boards and AI literacy programs. On the other hand, the responses seemed to vary in terms of the risk scope (systemic society-wide vs application-specific concerns), documentation methods, and depth and maturity of the AI safety research within an organization.



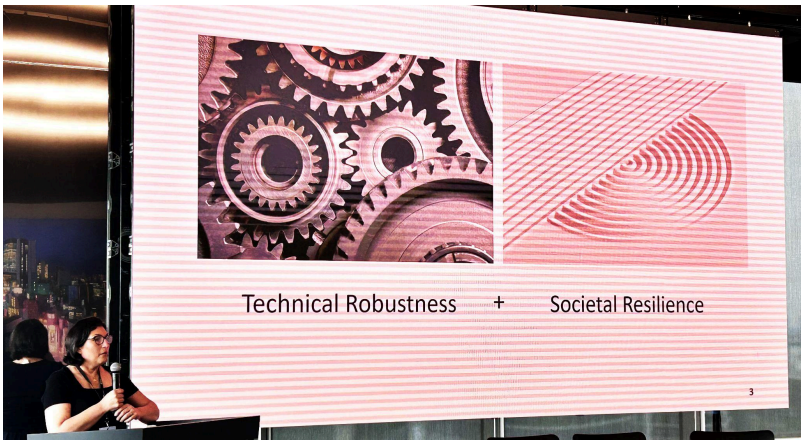
Additionally, Perset shared that for some companies HAIP provided an opportunity for the organization and researchers to "practice" reporting. Some organizations found the exercise very useful internally to understand gaps in their governance and/or documentation. As per the presentation, there were also some interesting differences in how various themes are approached depending on resources. For example, some large companies were using AI-based tools to identify AI risks, as well as real-time monitoring and (especially) red teaming. In terms of transparency, consumer-facing businesses seemed to prefer publishing model cards, while B2B organizations preferred sharing such information under bilateral confidentiality contracts.

Following the OECD's recorded presentation, one of the key architects of HAIP – **Yoichi Iida** from the Ministry of Internal Affairs and Communications of Japan – provided a short commentary. He shared some history on the development of the OECD AI Principles starting from 2016, which served as a basis for the 2023 Hiroshima AI Process. He noted that HAIP focuses on mechanisms rather than values. From him, participants received an update regarding the next steps for the Framework. Iida underlined the importance of voluntary collaboration with these companies, and how the reports need to provide meaningful information to advance transparency and interoperability. He noted that the next steps will focus on how to improve or streamline the process, especially for those organizations with less resources. He also highlighted the necessity to create guidelines for AI users to better understand and utilize these reports.

Lightning Talks

These initial opening remarks were followed by 4 short lightning talks about the current state of AI safety evaluations (technical and sociotechnical), the challenges, and what should lay ahead. The presentations were followed by a short round of Q&A for all the panelists.

The first presentation was delivered by **Elham Tabassi**, Director of the Brookings Artificial Intelligence and Emerging Technology (AIET) Initiative, and former Chief AI advisor at NIST, leading its AI Innovation Lab. Elham first shared experiences from her time at NIST and the successful publication of the NIST AI Risk Management Framework, and the importance of the science of evaluations. She remarked that the "question is not 'Is AI safe?' but 'How does safety emerge in practice?'" Focusing on the need to measure real world complexities, she highlighted that we need to consider technical capabilities alongside the deployment context and the human behavior. While the models can be evaluated for capabilities, such evaluations need to be accompanied by stress testing ("examining risks that may materialize") and field testing ("evaluating real-world context of



use”). Emphasis should be on technical robustness and societal resilience in order to ensure AI benefits everyone. In other words, it should not be just about “whether we need AI safety evaluation – it’s whether we can develop frameworks adaptable and sophisticated enough to handle AI’s integration into society.” Tabassi warned that there is currently an overreliance on laboratory testing, while deployment context integration is poor, and continuous monitoring is lacking. She warned that the field currently does not have a clear, concise understanding of what to measure. "What does good look like?" What do we mean by "trustworthy?" What do we mean by "safe?" Most of the evaluations are still preliminary and dominated by a few actors. Tabassi concluded that we need more science-based methods and metrics, community participation, and incentives for companies to share and report their evaluations.

The second lightning presentation was a recorded one from **John Burden** – Cambridge University – Leverhulme Centre for the Future of Intelligence. John’s research focuses on the challenges of evaluating the capability and generality of AI systems, how these concepts relate to risk posed by the system and how such risks can be mitigated. John noted that although we are doing science, we need to work on science for the future, science that takes into account future developments. His presentation focused on evaluating the safety properties of general-purpose AI (GPAI) which can perform a wide range of tasks and can be easily accessed by the public. John differentiated between dangerous capability evaluations (what is a system capable of doing), and alignment evaluations (what a system will try to do, regardless of its capabilities). Additionally, he listed other safety concerns such as bias, the need for interpretability in high-risk domains, robustness against adversarial attacks and accidental failures, and AI systems contributing to dangerous capability uplift.

He noted that some of the most widely used evaluations such as benchmarking or red-teaming have significant limitations. In the case of benchmarking, John listed some of the major issues as:

- lack of construct validity: where benchmarks do not necessarily measure what they claim to,
- target for optimization: referencing Goodhart’s Law where companies are more focused on achieving leadership in the benchmarks and higher scores, but not necessarily the meaningful use of the models,
- contamination of training data of benchmarks: due to leakage into training data, which the performance levels are inflated,
- limitations in variety: where benchmarks cannot possibly cover every scenario or variation of questions, and
- low quality training data: which is included in the benchmarking sets resulting in inaccurate measurements.

As for the red-teaming, John listed the major limitations as:

- limitations in variety and quality of effort: where the red-teamers are either not able to ask all possible questions / scenarios where the system could be unsafe,
- false assurance: in cases where the red-team prompts may not surface anything problematic but then 'absence of evidence is not evidence of absence,'
- Cycle of testing: that is required to conduct new red team experiments after the system is fine-tuned in response to a previous red team experiment

The next lightning talk was from **Kit Kitamura**, Principal Expert at Japan AI Safety Institute and Visiting Researcher at the Institute for Future Initiatives, The University of Tokyo and at the National Institute of Advanced Industrial Science and Technology (AIST). Kitamura was also one of the main organizers of this AI Safety Workshop, contributing many hours with his colleagues from AISI and AIST. He opened his presentation with a reminder that innovation and safety should be pursued simultaneously. This requires a proactive approach to safety, and concrete integration into corporate, product, and service strategies. Rather than reacting passively, Kitamura said "we must anticipate and lead technological and socio-technical inflection points in AI ecosystems and social systems." He proposed introducing a dialogue-based culture in Japan to complement the high degree of social discipline and collective responsibility. Such deliberations are critical, and this Workshop series indeed envisages dialogue-based approach to community building and AI safety.

The final distinguished speaker was **Yutaka Oiwa**, the Deputy Director of Intelligent Platform Research Institute (IPRI); Project Leader for AI safety research project / J-AISI partnership and the National Institute of Advanced Industrial Science and Technology (AIST). Oiwa provided a historical timeline of AIST's work on AI safety and machine learning quality management. He warned about the urgency of action by demonstrating the narrowing time gap between development of a technology and its implementation in critical ways. Whereas traditional software had a 5 to 10 year gap for meaningful deployment scale, Large Language Models (LLMs) shortened that gap significantly. This means that our response times need to match the speed of technological progress. Oiwa suggested that as much as we can, we should apply forecast-based, proactive AI safety so that we can be more ready with safety responses at or before the point of real-world deployment of new technologies. A current example of this could be thinking about guardrails for possible safety risks introduced by (multi)agentic AI, or agent-human-agent interactions. He also highlighted the importance of evaluation of risks and safety within context, and suggested that sandboxes, problem challenges and similar activities can help move the needle in the right direction.

Breakout Discussions

The second half of the AI Safety Workshop was the interactive session with all the participants. Participants were split into 4 groups (2 x Technical and 2 x Sociotechnical) and were asked to discuss the following questions. The conversation in these groups was skillfully facilitated by researchers from J-AISI and AIST. As far as the organizers are aware, this session format for an AI safety event was also the first of its kind in Japan.

The diversity of the participants here was critical as their experience, knowledge and variety in perspectives allowed them to tackle the issue from different angles and challenge their own assumptions.

- What should be addressed regarding AI safety risks (technical and sociotechnical)?
- How should these AI safety risks (technical and sociotechnical) be addressed?
- Challenges in applying AI safety mitigations?
- Discuss potential alignment across international efforts. Are there regional different elements or meanings of safety approaches?

At the end of the one-hour breakout discussions, each group was asked to present their discussions. The following provides a summary of these deliberations.

What Should Be Addressed Regarding AI Safety Risks? (Technical and Sociotechnical)

A nuanced understanding of AI safety risks requires that we identify both the technical vulnerabilities within AI systems and the broader sociotechnical context in which these technologies operate. Such understanding also helps organizations develop more informed strategies for AI adoption and targeted investments for AI safety, governance and quality management. The AI Safety Workshop's intention with this first question was to lay out the safety risks participants deemed most important. In other words, map out the risks participants were most concerned about with regards to their practices. There is an obvious difference in responses to this question between the research community and those looking to adopt and implement AI in their respective organizations. However, participants agreed on the need for greater precision in defining and classifying risks. Vague discussion around "risk" is inadequate; policies and governance must determine both acceptable risk levels and the trade-offs society is willing to make. This typology must account for a spectrum of harms, from unintended bugs and accidents to deliberate, high-stakes misuses, and ensure a focus on the highest risks first, especially those with the greatest potential to threaten human life or key societal functions.

Expanding Capability, Expanding Risk

Participants recognized a direct relationship between increasing AI capabilities and the proliferation of risks. The emergence of more complex and flexible systems - multimodal models, agentic AIs, multi-agent interactions, autonomous decision-making - unlocks a broader range of harms, both deliberate and accidental. Concerns ranged from old threats, like cyber misuse and criminal exploitation, to frontier risks, like autonomous weaponization or time-delayed harmful actions (the so-called "time bomb" scenario). The ability to engage with general-purpose AI systems simply using AI has significantly reduced the level of technical knowledge and experience which was previously required for malicious purposes. At the same time, given the unconstrained nature of language, the threat surface for attacks has expanded exponentially. Humans are creative and their creativity with language can be seen in many prompt injection examples.

Dangerous Capabilities and Misalignment

A recurring theme was the threat posed by advanced capabilities beyond human oversight. The issue of "alignment"—whether AI objectives are reliably constrained to be compatible with

human values and societal well-being—remains unsolved, and potentially existential. While the current AI technologies are not able to autonomously pursue goals, nevertheless misalignment can introduce significant risks when deployed in real-world environments. Misalignment and optimization problems, first experienced in predictive AI systems, continue to be problematic with general-purpose AI systems. In fact, the general-purpose nature of these systems makes it even more difficult to hard-code fail-proof constraints. With the integration of natural language into agentic systems as well as humanoid robots, the concerns for misalignment remain high.

Building upon the theme of malicious intent, participants were also concerned that current systems may be repurposed for dangerous outcomes, such as creating biochemical weapons, performing cyber-attacks, or engaging in sophisticated influence operations. The systems currently do not engage in these dangerous activities autonomously, but they can contribute or uplift the ability of malicious actors to do so.

Structural and Infrastructural Risks

Other friction points included the speed of technological evolution, compounded by insufficient field testing. While many of the known issues with AI systems such as (privacy, explainability, bias) are not yet meaningfully addressed, technology is becoming even more complex. Ability for stakeholders (researchers, corporate implementers, policymakers, regulators) to respond to the evolution and develop timely evaluations and guardrails is hampered. As per the participants, some of the issues seemed to be due to the opacity of foundation models (large, general-purpose models upon which many applications are built). These black-box systems stymie efforts at predictability, accountability, and safety verification. Coupled with unequal access to compute and data and evaluation processes monopolized by a few commercial actors, systemic issues arise that go far beyond the technologies themselves.

Participants noted that such monopolization reduces the incentives for these actors to invest more into AI safety, or to evaluate their products thoroughly before deployment. Another structural issue raised by participants echoed the earlier remarks in the lightning talks about the limitations of benchmarks. Only a handful actors are involved in the development of such benchmarks - which may ultimately serve their own purposes, or simply not reflect real-world context. One participant noted that companies could also possibly understate the AI's actual capability through sandbagging methods.

Challenges of Human Factors

A further insight was the fundamental problem that developers are often the least reliable testers of the systems they build. Their mental models are constrained by how they think

systems “should” be used, often missing how they will “actually be used” —and misused—in practice. This human element introduces an additional layer of unpredictability.

Similarly, the unpredictable or “stochastic” behavior of AI complicates assessments—models may fail in ways their designers cannot anticipate.

Societal Resilience and Human Dependence

Beyond immediate technical harm, participants underscored broader social resilience risks. Growing dependence on AI (e.g., companions, vulnerability to disinformation due to consumption of information only through social media), coupled with the opacity of how AI models operate, could undermine human autonomy, societal trust, and the functioning of critical infrastructures. The demand for simultaneous evaluation of both technical robustness and societal risks was clear. It is not enough for AI to function as intended; but it must do so without causing unpredictable harm in complex real-world environments.

Participants were enthusiastic about AI augmenting, supercharging, or even automating some tasks and processes. However, a need for ‘justified trust’ was repeated.

How Should These AI Safety Risks (Technical and Sociotechnical) Be Addressed?

Once the participants shared these risks which were most critical to them, they were asked for possible solutions. Participants offered a multiplicity of solutions, emphasizing that no single approach suffices; rather, an interdisciplinary, multi-layered framework is essential to making progress on AI safety.

There was consensus across the workshop that AI systems should enable safety, where it is (1) easy to do the right thing with the system, (2) difficult to do the wrong thing, and (3) easy to recover when the wrong things do happen.

Participants emphasized the importance of foundational characteristics for safe AI. These foundational characteristics stood out as prerequisites for trustworthy AI, including:

- Validity and reliability - ensuring AI systems perform as intended.
- Security, robustness, and interpretability - so that their actions can be understood, challenged, and improved.
- Harmful bias mitigation and privacy enhancement - mitigating harm as much as possible and protecting sensitive data.

- Contextual tailoring - as risk varies widely by domain (e.g., battlefield AI vs. medical diagnostics) and the characteristics should be tailored to domain of application.

Comprehensive, Contextual Risk Models

Attempts to manage AI safety must rely on comprehensive, domain-specific risk models. Participants noted that the safety risks span a **spectrum** - from **narrow issues** like bias and technical robustness to **existential risks**. Similarly, AI safety concerns differ widely by domain (e.g., autonomous vehicles vs. military vs. science). Different scopes require different mindsets—regulatory, ethical, technical, or geopolitical—and no single framework can comprehensively govern all. These need to be built upon the foundational characteristics for trustworthy AI though.

Participants noted that while AI brings certain novel risks and challenges, there are still many learning opportunities from adjacent fields (aviation, medicine, automative, civic engineering, nuclear safety) to translate best safety practices for the AI context.

Another concern was the interconnectedness of the components of AI systems, as well as interconnectedness between AI systems or agents. While individual mitigations matter, risks often emerge from the interaction of many elements within a system. Systems are more complex than their individual parts. These conversations surfaced a real need for systems level assessment and governance in the future.

Testing, Monitoring, and Benchmarking

Laboratory testing and field deployment must be complemented by ongoing monitoring and meaningful benchmarking. A key question is how to capture the earliest signs of risk development and intervene before harms actualize or spread. Methods such as red-teaming seem to be adopted widely due to lower cost and resource commitments. However, participants were concerned that red-teaming may create a false sense of assurance despite its significant limitations. Complementing model-centric evaluation with context-of-use–centric evaluations could surface additional risks or those emerging from real-world deployments.

Capability-oriented evaluation is a new approach that focuses on making AI behavior more predictable. However, information sharing with stakeholders outside the corporate labs still remains limited. Collaborative efforts to partner with academia and civil society for pre-deployment safety evaluations could significantly improve the current practices. Participants noted that such collaborations also need to extend beyond borders because

deployment may look very different across cultures, geography and industries. Objective, public interest benchmarking seemed to emerge as a possible alternative to the current methods.

Additionally, robust incident sharing mechanisms were deemed to be helpful to both researchers and corporate implementers. Such incident reporting provides wider awareness of issues and more effective responses. Aviation industry's established methods of accident/incident reporting, analysis, dissemination of lessons learned for the industry was cited as a best practice. Some participants noted that such transparency, monitoring and collaboration could lead to more trust (into the process, actors, AI tools) despite the incidents.

Separation of Foundation Models and Applications

Participants noted that one way to address safety challenges could be separating the foundational elements of AI models from the many applications that use them. This division clarifies where different types of responsibility lie and may help prioritize mitigations to those players with the greatest leverage: the foundational model developers.

Some of the technical mitigations which were suggested included neuro-symbolic models and formal methods for enforceable output constraints, domain-specific specifications and filtering, and use of auditing mechanisms, including activation analysis and sandbagging detection (i.e., identifying when models alter behavior during evaluation).

Standardization and Best Practices

Strong calls emerged around developing engineering standards (such as domain-specific "gold standard" protocols and nascent industry efforts like the Model-Context-Protocol, MCP).

Participants noted that the manual safety evaluations (e.g., via Scale.ai) are expensive and dominated by a few companies. Engineering standards and open-source evaluations could help democratize the practices.

Calls for engineering standards were accompanied by a need to have unified guidelines for the bare minimum before we can start domain-specific implementations. Particularly, danger to human life must be averted and this should constitute a bare minimum.

Accountability, Liability, and Distribution of Responsibility

Shared responsibility emerged as a paramount need. Liability for unsafe outcomes must extend up and down the pipeline: from foundational model creators to application developers, deployers, infrastructure providers, and even end-users. Model legal frameworks such as those

used in transportation (e.g., car accidents) may provide useful analogies. One recommendation was to require stakeholder listings from developers.

Participants noted that carrot and stick (incentives / recognition / prizes / contracts versus fines / sanctions / prohibitions) methods vary as per culture too. East Asian countries seem to prefer observation and balancing of actual behavior and harms first, followed by guidance. This is in contrast to the pre-cautionary approach of Europe, and reliance on regulators and courts in the United States.

Both unintended harms (bugs, accidents) and deliberate harmful uses (e.g., autonomous weapons) must be considered within AI safety scope. Nevertheless, there was agreement that user error or intentional misuse are not the only sources of risk and that the frontier model companies carry the biggest responsibility. It was highlighted that these labs should be more forthcoming about their activities, decisions and supply chains. They are literally the "foundation" upon which others build. Risks should be detected and mitigated during early development.

The downstream businesses who build applications on top of the GPAI products carry significant dependence risk. The respective level of risk / responsibility of the involved entities should be estimated and then enumerated. Participants who are considering or working on implementing GPAI models in their organizations noted the impact of evaluations on their adoption decisions. More transparency about safety concerns, evaluations and mitigations from frontier model companies were highlighted as increasingly important.

AI Literacy, Training, and Industry-Wide Testing

A lack of AI literacy—among both developers and users—was highlighted as a persistent challenge. Calls for widespread mandatory literacy programs, certification-like regimes, and broad stakeholder inclusion were repeated elements. Participants noted that the more informed and trained developers, implementers, and users are, then the more people are involved in detecting safety issues, surfacing use cases, and driving adoption.

Finally, mitigations must involve not only technologists but also stakeholders from civil society, regulatory authorities, and end-user communities. Open information sharing, participatory standard development, and public accountability were repeatedly emphasized. Technical methods alone are insufficient—the sociotechnical context must guide the application of every solution.

Challenges in Applying AI Safety Mitigations

Despite the clarity around potential AI safety concerns and possible solutions, the workshop responses reflect many challenges to effectively implementing AI safety measures.

Ambiguities in Definitions and Risk Boundaries

Participants noted that the concept of 'safety' currently means many different things to many people. This not only creates confusion for practitioners but also undermines genuine efforts to evaluate and govern AI systems for safety concerns. Additionally, uncertainty remains around what constitutes "sufficient" safety or the "right budget." Several participants emphasized that some capabilities—recursive self-improvement, biological weapons development, autonomous self-replication—should simply be outlawed due to catastrophic potential. Still, there was no consensus on how to systematically prevent their emergence.

Similarly, law needs to determine the level of risk that is acceptable (and unacceptable). Although the risk can never be brought to an absolute zero, participants noted that there must be a clear definition and agreement of what is unacceptable.

Velocity of Progress Outpasses Oversight

The speed of AI progress, especially with foundation models, was again mentioned as a significant challenge to applying AI safety mitigations. It continues to outstrip the ability of regulatory, safety, and research communities to keep pace. This creates a constant game of catch-up where risks appear and evolve faster than safeguards can be meaningfully applied. While it is natural that these components follow technological progress, the gap might be widening.

The Evaluation Bottleneck

Manual safety evaluations are expensive, slow, and frequently monopolized by a handful of industry actors, leaving most of the ecosystem underserved. Small organizations, in particular, lack budget and expertise for meaningful participation. Even the transparency technologies, such as watermarking and metadata signing, are mostly limited to major companies, with adoption remaining narrow in scope.

Similarly, efforts to "tick boxes" for safety can become superficial "safety washing," where compliance exists in name only.

Both the presenters and the participants highlighted that many evaluation methods are still preliminary and lack a strong scientific basis.

Weak Incentives and Insufficient Funding

Participants highlighted that the financial and organizational incentives are poorly aligned with safety – mostly across the major tech companies. Currently, only about 1% of AI research budgets are allocated to safety, far short of investments in comparably high-stakes domains like aviation or nuclear power (where up to 90% of resources focus on safety and compliance). The differences in industry practice here could help guide AI policy, e.g., dedicating specific budget percentages to safety, certification, ongoing audits, and mandatory education. Participants argued that 10-20% should be the minimum threshold for acceptability. Moreover, the budget for safety must include not only direct costs but also the compute resources to perform robust testing and verification.

Another challenge concerned the possibility or risk of overfitting evaluation benchmarks for marketing rather than real safety.

Lack of Transparency and Collaboration Gaps

Opaque models, particularly “black box” foundation systems, make external inquiries and accountability hard.

Insufficient cross-sector collaboration—especially between industry and research community (including civil society) - limits the depth and validity of governance models. Major technology companies can be biased toward economic or strategic goals, downplaying or delaying inconvenient safety measures. However, this creates technological debt and also additional burdens and risks on the downstream businesses, individuals and communities. While participants understood that certain risks cannot be eliminated, they also noted that a minimum level of transparency about data, models and design decisions were necessary for them to properly assess their risk tolerance and put counter measures in place.

Alignment Across International Efforts: Regional Differences and Pathways to Coordination

The aspiration for aligned, international efforts on AI safety must reckon with natural regional differences, both in values and practical priorities. Countries and regions may differ in terms of

culture, language, social system, political system, demographics, and economical situation. Such differences may shape how “safety” is defined and prioritized.

For instance, participants noted that Asian countries, such as Japan, focus on social implementation and transformation, trustworthiness, and the development of governance frameworks, while European and North American approaches frequently stress transparency, accountability, and interpretability in public-facing domains.

At the same time, socio-technical perspectives such as interpretability and social legitimacy are particularly emphasized in public domains such as finance, education, and healthcare, and are therefore considered to be highly compatible with accountability-oriented approaches seen especially in Europe and Asia.

Participants agreed that collaboration across organizations and national borders is a key factor for success.

Flexible, Context-Aware Frameworks

While there is broad agreement on the abstract goals of AI safety, governance should be sufficiently flexible to account for institutional, cultural, and societal variation. This means respecting local values and not imposing one-size-fits-all solutions. Socio-technical aspects, such as ensuring systems are legitimate in the eyes of local populations, are crucial to achieving traction and trust. Thus, frameworks must balance a common core (e.g., fundamental rights, basic safety standards) with local adaptation, especially for high-impact sectors like finance, healthcare, and education.

Strategic Levers for Effective Global Collaboration

Key recommendations for enhancing alignment included:

- Promoting inclusive, transparent standardization, as well as transparency about who works on safety-critical models,
- Tracking talent and compute concentrations, as these remain key chokepoints for power and capability in AI development,
- Building and democratizing AI safety evaluation tools and auditing infrastructure,
- Deploying public-facing AI incident repositories so that others do not repeat known issues, and researchers can contribute to safer approaches,
- Distributing responsibility along the entire development and deployment pipeline,
- Elevating AI literacy and awareness among stakeholders, embedding training into workplace practices and public policy,

Conclusion and Next Steps

AI safety is perhaps the archetypal “wicked problem”: deeply technical, highly social, and permanently in flux. The core themes emerging from workshop participants are the need for ongoing discussion, collaboration, and evaluation. Consensus exists around core risks and demand for better and more scientific safety evaluations, clearer communication about the safety practices, and the need for greater budget and accountability.

True progress will require not only cross-organizational and international collaboration but new forms of governance that can keep pace with the industry’s leading edge.

As noted at the beginning, the AI Safety Workshop held on July 5th, 2025 was intended to be the first in a series of community building and engagement activities focused on AI safety, governance and quality management. With the series of AI safety workshops, and continuous discussions in between – the objective is to help build and strengthen a community of AI safety professionals and practitioners and hopefully bridge some of the existing efforts in Japan and beyond.

The next steps include 1)post-event survey to evaluate participants feedback on the quality of the first event, as well as topics for future ones, 2)launch of a Slack channel to include the participants of the workshops. The channel will help efforts towards community building and engagement activities, creating an inclusive space to ask questions, share resources, connect with others, invite others to other relevant gatherings/events, suggest other practitioners / researchers for future discussions.

Acknowledgements

The AI Safety Workshop was held under Chatham House rules to encourage deeper conversations. Therefore, the participant names are kept confidential. However, many individuals and organizations contributed to the planning and execution of this event and we would like to express our sincere gratitude to the following for their insights, time and passion.

Organizations (alphabetically):

Chiba Tech University – Henkaku Center, Digital Garage, Japan AI Safety Institute, Organisation for Economic Co-operation and Development (OECD), The National Institute of Advanced Industrial Science and Technology (AIST)

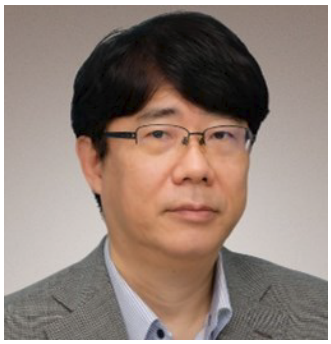
Individuals (alphabetically):

- Akiko Murakami (Japan AI Safety Institute)
- Daum Kim (Chiba Tech University – Henkaku Center)
- Elham Tabassi (Brookings Institution)
- Fiza Razik (Chiba Tech University – Henkaku Center)
- Hikaru Matsuoka (Japan AI Safety Institute | RIKEN)
- Hiromu (Kit) Kitamura (Japan AI Safety Institute)
- John Burden (Cambridge University – Leverhulme Centre for the Future of Intelligence)
- Joichi Ito (Chiba Tech University – Henkaku Center)
- Joseph Park (Digital Architecture Lab)
- Karine Perset (OECD)
- Kazuaki Nimura (Japan AI Safety Institute)
- Kazuhiro Taga (Japan AI Safety Institute)
- Keita Azuma (Japan AI Safety Institute)
- Kenji Hiramoto (Japan AI Safety Institute)
- Koichi Konishi (National Institute of Advanced Industrial Science and Technology)
- Lulu Ito (Digital Architecture Lab)
- Merve Hickok (Chiba Tech University – Henkaku Center)
- Mizuki Oka (Chiba Tech University – Henkaku Center)
- Samanta Bates (Chiba Tech University – Henkaku Center)
- Shin Nakajima (National Institute of Advanced Industrial Science and Technology(AIST), National Institute of Informatics (NII))
- Takehito Akima (Japan AI Safety Institute)
- Tim Burress (Digital Garage)
- Yoichi Iida (Ministry of Internal Affairs and Communications of Japan)
- Yoshiki Seo (National Institute of Advanced Industrial Science and Technology)
- Yuki Yamamoto (Japan AI Safety Institute)
- Yuma Kurihara (Japan AI Safety Institute)
- Yutaka OIWA (National Institute of Advanced Industrial Science and Technology)

And all the facilitators, note-takers and presenters who voluntarily contributed to the success of the Workshop.

Merve Hickok - Visiting Researcher - Henkaku Center

Samantha Bates - Visiting Researcher - Henkaku Center



Hiromu (Kit) Kitamura - Principal Expert for Technical Management (in Charge of Supervising and Driving Technical Initiatives) - Japan AI Safety Institute

Joseph Park - Content Lead - Digital Architecture Lab