NEXT Alliance Japan Study trip: Tokyo, 2025

Allied on AI Assurance: Technology and Policy to Enhance AI Safety and National Security

Study Trip Report: Published April 2026

James L. Schoff
Distinguished Senior Fellow (Non-Resident)
Sasakawa Peace Foundation USA

Allied on AI Assurance: Technology and Policy to Enhance AI Security

Winter 2025 NEXT Alliance Study Trip, summary by James L. Schoff, Sasakawa USA

Preface

Sasakawa Peace Foundation USA, through its US-Japan NEXT Alliance Initiative (NEXT), organized a second "AI Assurance Study Trip" to Tokyo in December 2025, featuring 6 US experts in fields of Artificial Intelligence (AI) (especially AI Security and Safety) and international science collaboration. The purpose of this Study Trip was to foster networking, share information, and explore opportunities for bilateral technical research collaboration to help strengthen AI defenses, bolster cybersecurity, improve data quality, and help mitigate disinformation, all as a bilateral supplement to multilateral efforts.

The Sasakawa USA-funded Study Trip supports NEXT's project "Allied on AI Assurance: Technology and Policy to Enhance AI Security." NEXT organized the trip in collaboration with Racke Strategies & Technologies, as part of a "Japan-US Strategic Science, Technology & Innovation Initiative" (JUSSTII) activity, and with organizational support from Japanese partners.[1] We brought a diverse group of US AI security experts to Tokyo for 3 days of meetings/roundtable discussions with Japanese counterparts (government, academia, private sector). In addition, together with the University of Tokyo, Research Center for Advanced Science and Technology (RCAST)/Economic Security Intelligence Lab (ESIL) we co-hosted a public event attended by about 110 guests made up of students, researchers, private sector representatives and various embassy officials.

The US delegation participants included:

- *Mr. Matt Chessen, Resident Technical Expert at RAND's Center on the Geopolitics of Artificial General Intelligence*
- *Ms. Marissa Dotter, Lead AI Engineer, Trustworthy & Secure AI, MITRE Corporation*
- *Mr. Michael Goldman, Division Leader for the Computational Engineering Division; Lawrence Livermore National Laboratory*
- *Dr. Nidhi Kalra, Sr. Information Scientist at RAND; Prof. of Policy Analysis at RAND School of Public Policy*
- *Dr. Dimitri Kusnezov, Vice President for Science and Technology, Nuclear Threat Initiative and former Under Secretary for Science & Technology, Dept of Homeland Security*
- *Mr. Douglas Rake, President and CEO; Racke Strategies & Technologies, Inc.*
- *Mr. Jim Schoff, Sr. Director, NEXT Alliance Initiative, Sasakawa Peace Foundation USA*

Table of Contents:

1. Background and Introduction
2. AI Assurance Developments Since the First Tokyo AI Study Trip
3. AI and Cybersecurity
4. Physical AI Security
5. The US-Japan Alliance and AI Assurance

[1] JUSSTII is an intermittent series of Track 1.5 bilateral dialogues with a mission of promoting collaborative research in cutting edge science and technology areas of mutual strategic importance to the US and Japan. Sasakawa USA is grateful to Japan's National Institute of Information and Communications Technology (NICT) and Prof. Akira Igata for their organizational support with some of our events in Tokyo.

1. Background and Introduction

A phenomenon some call "adversarial AI" has become an economic and national security concern not only in terms how it can threaten cybersecurity efforts or AI training model integrity, but also its potential to spread disinformation and ultimately dampen AI utilization. As the 2025 US AI Action Plan notes, the inner workings of frontier AI systems are poorly understood, lack predictability, and are susceptible to adversarial inputs in ways that can undermine national security. Cybersecurity is a particular area of concern, where the relative strengths of AI for offense versus defense are still in question. In addition, physical AI security (i.e., for AI systems with integrated mechanical components that perform physical activities with real-world sensor feedback) is of growing interest amid accelerated incorporation with autonomous systems and robotics. At the same time, most nations are trying to avoid regulating AI development in ways that inhibit innovation. The US Trump administration, in particular, has been adamant that it will approach AI "not with fear, but with responsible boldness" and that it will "push back against global AI governance at the UN, G7, APEC, and other forums to defense great American companies from foreign nations' stifling regulatory regimes."[2] Japan has vowed to become "the most AI-friendly country in the world."

The US and Japan are wrestling with these challenges individually and collectively through multilateral forums such as the G7 and OECD, but bilateral networking on this topic can be an important supplement, especially for the technology behind policy development. Bilateral collaboration in AI fields is emerging through such channels as the Cross-Pacific AI Initiative (X-PAI) involving US and Japanese universities and companies, a similar initiative involving Carnegie Mellon University and Keio University funded by Arm and SoftBank Group, and Japan's National Institute of Information and Communications Technology's (NICT) CREATE program supporting cooperation with US firms and universities, among others. Our NEXT Study Trip aimed to build upon these initiatives by sharing best practices, comparing and aligning strategies, and exploring potential cooperative projects or investments, as Japan was finalizing its AI Basic Plan by the end of 2025. Bilateral collaboration in confronting AI-enabled disinformation and adversarial AI can be a force multiplier for investing in research to develop countermeasures and mitigation strategies, but a necessary first step is to enhance mutual understanding on the topic and grow the human network.

Each member of our US study trip delegation is involved in AI security and research or policy making in some way, either conducting government-sponsored research, supporting government policy makers with technical expertise, or facilitating bilateral science and technology collaboration in AI and related fields. The group's itinerary in Japan involved a series of not-for-attribution meetings and dialogues with Japanese counterparts designed to exchange information about current work,

[2] Michael J.K. Kratsios, Director, White House Office of Science and Technology Policy, "Statement before the Subcommittee on Research and Technology, Committee on Science, Space, and Technology," US House of Representatives, January 14, 2026.

share perspectives about challenging aspects of AI safety, and explore opportunities for bilateral collaboration. We are grateful to all the officials and experts who took the time to talk with us.

On Day One of the trip, the US delegation carried out a meeting with AI and economic security-related officials at Japan's National Security Secretariat, followed by discussions with a broad group of US Embassy experts and stakeholders across fields of technology, the economy, and national security. Sasakawa USA also organized a two-hour workshop devoted to physical AI security involving university AI researchers, government officials, and private sector representatives sharing perspectives about physical AI security concepts and research priorities. A second session explored possible opportunities for bilateral collaboration in these areas.

On Day Two, Sasakawa USA co-organized with Japan's National Institute of Information and Communications Technology (NICT) a three-hour workshop on mitigating cyber misuse risks for AI systems featuring national laboratory and university AI researchers sharing examples of current work and discussing collaboration strategies. The US delegation also made a site visit to Japan's AI Safety Institute and met with experts there.

On Day Three, the US group was hosted for an extended roundtable discussion by experts in Japan's Cabinet Office, Secretariat of Science, Technology and Innovation Policy (AI Office). The meeting included officials from Japan's Ministry of Education, Culture, Sports, Science and Technology (MEXT); Ministry of Economy, Trade & Industry (METI); and the Ministry of Internal Affairs and Communications (MIC). Earlier in the day, some members of the US delegation participated in the public event referenced in the Preface.

The following report combines insights and information gleaned from the study trip with subsequent archival research to update our 2024 report and dive deeper into areas of physical AI security and AI related to cybersecurity. It concludes with observations regarding opportunities and challenges for bilateral and "minilateral" collaboration on AI assurance.

2. AI Assurance Developments Since the First Tokyo AI Study Trip

Our 2024 project report included a quick review of AI-related terminology, and it is worth revisiting these labels and descriptions that are still becoming harmonized internationally and across different professional sectors. The US Commerce Department's National Institute for Standards and Technology (NIST) in 2024 began collaborating with Japan's new AI Safety Institute to create a "crosswalk" (i.e., a mapping of concepts and terms) of the NIST AI Risk Management Framework and Japan's AI Guidelines for Business. Crosswalk 1 of this process focused on terminology and has been produced in both English and Japanese. The overall aim was to promote interoperability between US and Japanese governance frameworks.

Evolving AI assurance concepts and governance initiatives

Within the broad topic of AI governance lies a concept of **Assurance**, what the MITRE Corporation in the US describes as "a process for discovering, assessing, and managing risk throughout the life cycle of an AI-enabled system so it operates to the benefit of its stakeholders." The US Department of Transportation describes AI assurance as "the techniques, activities, and processes used to evaluate and assure expected properties of AI components or AI-enabled systems" throughout their lifecycle. In its February 2024 report "Introduction to AI Assurance," the UK's Department of Science, Innovation and Technology described it as the measurement, evaluation and communication of the "trustworthiness of AI systems" based on credible evidence. **Trustworthy AI** is an important issue not only for consumers and businesses engaged with AI, but also in terms of public safety and national security. This latter area (national security) is where our NEXT "Allied on AI Assurance" project is focused.

AI Safety and AI Security are directly connected to trustworthiness and the tools necessary to evaluate them. According to the US-Japan crosswalk, the two countries define AI safety and AI security slightly differently. For NIST, AI is "safe" if it does not "lead to a state in which human life, health, property, or the environment is endangered [under defined conditions]." In contrast, Japan adds a "human-centric" component including consideration for human life, body property and mind, and the environment, as well as human "safety" from disinformation.

AI Security is also linked to trustworthiness (according to both countries), as it describes AI systems "that can maintain confidentiality, integrity, and availability" through reasonable protection mechanisms that prevent unauthorized access and use. This includes protecting the development of large language models (LLMs) from data "poisoning" or tampering, designing AI systems based on early and thorough adversarial AI risk assessments, and providing sufficient cybersecurity for AI systems and services throughout their lifespan.

Interestingly, Japan's AI Guidelines for Business add a note specifically mentioning that it is not possible to eliminate all vulnerabilities in AI systems and services, perhaps in part to reassure businesses that they won't be subject to unreasonable liability claims if something goes wrong. At the same time, the guidelines push companies to keep up with the "current technological level" to maximize safety. This admission of vulnerability points to another aspect of AI Security, namely mitigation. If adversarial AI is at some point destined to spark a crisis, then nations and governments will need tools available to help contain and reverse the damage.

There are several multilateral initiatives focused on enhancing AI safety and security. Notable among them is work carried out by the OECD through its Expert Group on AI Risk & Accountability and an Expert Group on AI Incidents, the latter of which created an automated monitor of AI incidents and hazards that now shows a monthly (6-month moving) average as of March 2026 for AI incidents globally at around 535. In comparison, for years leading up to spring 2023 the same moving average peaked a little above 100 incidents per month, and for most of 2024 the figure

hovered a little above 200. This AI Incidents Monitor (AIM) is designed to help to show risk patterns and establish a collective understanding of AI hazards and their multifaceted nature, so that it can be a useful tool for trustworthy AI.

The G7 has been another important multilateral forum to enhance AI assurance, often working in coordination with the OECD. In July 2024, for example, the OECD announced a pilot program to test a reporting framework linked to the G7's International Code of Conduct for Organizations Developing Advanced AI Systems. The program gathers information about how organizations developing advanced AI systems align with the Code of Conduct. In addition, the OECD and Global Partnership on AI (GPAI) formed an "integrated partnership" in July 2024 "bringing together all current OECD members and GPAI countries on equal footing, under the GPAI brand and on the basis of the OECD Recommendation on Artificial Intelligence." Japan's NICT established a GPAI Tokyo Expert Support Center in 2024.

The United Nations (UN) also contributes via its High-Level Advisory Body on AI and recently established (September 2025) "Global Dialogue on AI Governance," as do a series of global AI governance summits from 2023 (now involving over 80 nations in early 2026 in India) and various networks of AI governance-focused non-governmental organizations (NGOs) and industry groups. The UN effort, however, is being actively opposed by the United States and the Trump administration, which was one of only two countries to vote against appointment of a UN AI scientific assessment panel in February 2026 (117 countries supported the move). The panel's purpose is to produce regular assessments of AI's real-world effects across economic and social impacts and catalogue potential risks. Instead, the US Mission to the UN called it a "significant overreach of the UN's mandate and competence" and warned against any UN attempts "to regulate or stifle the development of cutting-edge technologies."

In addition, although the nationally-organized AI governance summits have grown in attention and attendance since the first UK "AI Safety Summit" in 2023, the most recent AI "Impact" Summit held in India carried on the recent trend of expanding the AI agenda while losing focus on more sensitive governance issues. Some new collaborative research initiatives focused on AI safety and cybersecurity were reportedly announced at the Summit, but with an estimated 300,000 attendees, it will be difficult for these grand international gatherings to advance specific actionable AI safety initiatives.

At a more practical and technical level, during a meeting with Japan's AI Safety Institute at our first AI Study Trip in 2024, a Japanese specialist highlighted the importance of "SC 42," which is a short-hand way of referring to a special subcommittee that began meeting in 2018 to help the technical side of international standard setting organizations address challenges unique to AI. SC 42 supports Joint Technical Committee 1 (JTC 1) that was established by the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) and has facilitated

standardization globally for such technologies as smart (chip/credit) cards, biometric systems, bar codes, cloud computing services, and many others.

The AI-focused JTC1/SC42 has helped publish 41 ISO standards related to AI, the most recent in 2025 focused on such areas as data quality governance, approaches for AI explainability and interpretability, and AI system testing. Japan contributes to SC 42 work and then Japan AISI benefits from SC 42 output to inform domestic regulation development and the design of AI safety testbeds. Other AI-relevant standard setting organizations include the International Telecommunications Union (ITU) and the Institute of Electrical and Electronics Engineers (IEEE).

The future of AI assurance is still likely to be multilateral, and in fact during one of the US delegation's meetings with US Embassy staff in 2024 an official said that the Embassy "thinks about what we can do with Japan multilaterally about 90 percent of the time. 10 percent is bilateral." But there is less conviction on this point in the United States under the Trump administration, so opportunities for bilateral coordination are growing relative to dwindling US attention on the multilateral front. One benefit of this role can still be the allies acting as a small caucus within broader multilateral frameworks for pooling information and building consensus that go beyond one nation's perspective (and set of interests) but are not as unwieldy as larger international groups. Another aspect is to combine strengths or compensate for weaknesses in terms of resources or expertise.

Although US firms and laboratories are world leaders in AI research and development, attracting nearly 80 percent of global AI research funding, Japan is still an engineering superpower providing value. Japan's AI skills in robotics and industrial systems in particular are world class, together with possessing a highly disciplined and rigorous academic research community for AI science and engineering. Japan is also a leader in information authentication technologies (e.g., so-called Originator Profile) that can help combat online disinformation. There are a variety of areas where US and Japanese firms and researchers can combine their strengths or complement each other to improve the speed and quality of their work as they contribute to multilateral initiatives or allied interests.

Related to this, the US and Japanese governments are investing significant sums of money in research and the development of tools to help them improve AI assurance for the sake of safe and secure AI systems. Their private sectors are investing even more, each in their own ways, and are valuable partners in this process. The US and Japanese governments are consistently numbers 1 and 3 in the world respectively in terms of the amount spent on science and engineering research and development. Japan has also been either the largest or second-largest foreign source of R&D investment in the United States since 2010, and it is the US's largest source of foreign direct investment.

It is also true, however, that various countries including Japan and the United States are paying greater attention to protecting or fostering certain national AI technology capabilities under a

broad banner of "AI sovereignty," which is a cross-current against collaboration. For the United States, this includes policies to protect some of its perceived advantages through export controls or investment restrictions, and it extends to promoting the export of a suite or "full stack" of AI capabilities to countries so they can avoid getting "locked into" a Chinese AI environment. For Japan, AI sovereignty includes collaborating with US initiatives, such as through a multilateral AI supply chain initiative dubbed Pax Silica, but also investing in its own domestic capabilities through subsidies, facilitating domestic access to data, and centralized compute and testing resources.

Japan and US governance approaches: diverging…slightly…

Following Japan's passage of a national AI Promotion Act in 2025, Japan's Cabinet Office created a special team to translate the broad Cabinet-approved strategy and published an "AI Basic Plan" on December 23, 2025.[3] The new law established an AI Strategic Headquarters chaired by the Prime Minister that is focused on promoting innovation to improve citizens' lives, support the national economy and national security, and enhance international competitiveness while also mitigating risks specific to AI. The Act applies to all persons, including non-Japanese, who engage in AI-related business activities in Japanese including research, development, and utilization.

Japan's AI Basic Plan tries to address two competing challenges: 1) to help the country catch-up to global AI powers by limiting development restrictions, fostering innovation, and becoming "the most AI-friendly country in the world," and 2) accelerating AI utilization in part by bolstering AI safety and "building a trustworthy AI ecosystem." For the former goal, Japan avoids introducing strict disclosure requirements for AI safety protocols and incidents, legally binding guidelines, extra-ordinary whistleblower protections, or damage liability specific to AI. At the same time, Japan aims to build a data linkage platform to support private R&D, invest in the AI workforce, and accelerate government adoption of AI while enhancing citizens' confidence in the technology. For the latter, Japan plans to double the size and budget of its AI Safety Institute, support AI incident reporting, and create non-binding guidelines for businesses utilizing AI in safe and transparent ways.

AI adoption rates in Japan are relatively low currently (at around 20 percent of Internet users, compared to 81 percent in China and 69 percent in the United States). A 2025 OECD report found that only 8.4 percent of workers in Japan reported using AI in the workplace, across all industries. The government hopes to support a virtuous cycle of demand-driven AI innovation in Japan that will boost productivity amid a declining population, support the economy, and reduce Japan's reliance on foreign providers of AI systems. Supporting Japanese trust in safe AI systems and products is a key part of this strategy to overcome a sense of wariness among the public, as only about half of polled Japanese believe AI offers more benefits than drawbacks. Americans are

[3] The official name of the act is "The Act on Promotion of Research and Development, and Utilization of AI-related Technology," passed in May 2025 and came into full force from September 2025.

among the world's most skeptical at less than 45 percent. Americans and Japanese are also among the world's least trustful that their governments can govern AI responsibly at around 33 percent.

The Trump administration also places a high priority on AI competitiveness and minimizing any possible impediments to AI development and entrepreneurship in the United States, but it has gone a step beyond the previous Biden administration to actively discourage state-level legislation that can be interpreted as going against Trump's policy "to sustain and enhance the United States' global AI dominance through a minimally burdensome national policy framework for AI." This prompted some push-back from both Democratic and Republican state leaders, and multiple states are taking legislative steps to promote AI transparency, protect children, address disinformation, and secure intellectual property rights and other measures related to AI safety, disregarding White House admonitions.

The first major state-level law enacted law in California in 2025, for example, goes further than Japan's national act in terms of a more rigorous reporting and oversight system, speedy incident disclosure requirements (subject to civil penalties), and enhanced whistleblower protections. As one think tank report noted, AI "transparency becomes a recurring obligation rather than an optional gesture," and each company's safety frameworks must be posted publicly rather than submitted privately to regulators. The new law does not go as far as an earlier version of the bill, however, that allowed for possible monetary and punitive damages related to "covered" AI models. That bill was vetoed by the California governor in 2024. New York also enacted new AI governance legislation in December 2025 that is similar to the California law, and Michigan is close behind.

Trump's AI Action Plan does mention the importance of safe and secure AI development, and it proposes building an AI evaluation ecosystem to help define and measure AI reliability. Subsequently, however, it tried to cut funding for AI safety research in 2026 by trimming support for the responsible organizations (i.e., National Institute of Standards and Technology (NIST) and its renamed AI Safety Institute).[4] House and Senate appropriators reinstated (and even slightly increased) the money for NIST and CAISI in January 2026.

The Trump administration's rejection of "global AI governance" (as noted in the "Background" section above) stands as another point of divergence from the Biden administration, and it represents a potential area of friction with Japan's approach. Japan's AI Basic Plan says that Japan will utilize all forms of international collaboration on AI including the Hiroshima [G7] AI Process and its "Friends Group," international standards bodies, the Global Partnership on AI and AI Safety Institute networks to help build consensus around AI governance norms. Japan stresses the importance of "strengthening cooperation with ASEAN and other Global South countries" in these areas as part of its international engagement plan.

[4] The Trump administration renamed the US AI Safety Institute as the Center for AI Standards and Innovation (CAISI) in June 2025.

Thus, although the US and Japan still share a similar desire for a relatively light regulatory touch for AI governance, in contrast to the EU and South Korea, their approaches to AI safety at the policy level are beginning to move in different directions on certain issues. However, as large US states adopt AI governance frameworks that are more in synch with Japan, an alliance consensus could emerge that facilitates coordination at the national level. Ultimately, it is likely that major safety AI incidents or discovered vulnerabilities will end up driving national debates around AI assurance, eliciting a policy response. In the meantime, collaborative research and dialogue among allies can help prepare us for AI safety and security challenges ahead.

3. AI and Cybersecurity

There are many ways that AI technology advances can both help and hinder efforts to manage cyber misuse risks for AI systems. Some analysts are concerned that these advances could empower cyber attackers more than defenders over time, making AI systems harder to protect precisely when their reach and capacities are peaking. The two-sided nature of this challenge was most recently revealed in April 2026 by a new cybersecurity tool from Anthropic called Claude Mythos Preview, which the company claims is so powerful that it is withholding public release until larger tech and financial firms have a chance to use it for protection first. Using this new product, Anthropic said that it found "thousands of high-severity vulnerabilities" in every major operating system and Web browser, and these could compromise critical infrastructure around the world. In response, Treasury Secretary Scott Bessent and Federal Reserve Chair Jay Powell called an emergency meeting of large bank CEOs to discuss possible implications of Mythos for the financial sector.

It is vital, therefore, to maximize allied abilities to identify and help manage AI cyber misuse risks in part through sharing best-practices, information, and resources, along with continuous improvement of techniques and technology. In partnership with Japan's NICT, our 2025 Study Trip members and other US and Japanese AI and cybersecurity experts met in Tokyo to share examples of current work and discuss possible collaboration strategies.

One NICT researcher described his recent work on source code analysis to identify defects, vulnerabilities, or malicious code, and then building an AI-powered source code analysis model that is reliable and robust. As an example, he noted a recent case of a malicious npm (node package manager) package using a hidden prompt to evade AI analysis tools. The challenge is even greater with the growth of AI-generated code (via LLMs) through so-called vibe coding and leveraging AI as an assistant or agent in this process. Research in 2025 showed that on average AI-generated code introduced risky security flaws in 45 percent of standard tests. The failure rate for one programming language was as high as 72 percent. This is roughly double the risk rate compared to human-generated code.

Another NICT researcher introduced his work on a slightly different problem, focusing on the explainability or interpretability of AI-powered cybersecurity intrusion detection systems (IDS). Where the first area of research aims to maximize confidence in the resiliency of soon-to-be published software, the second example is trying to improve the efficiency and effectiveness of cyber defenses after publication. Although modern IDS can be highly accurate at detecting cyber threats, they still generate a notable degree of false positives or misunderstood alerts, and without sufficient interpretability they can lead to wasted time, overlooked threats, and compliance risks. As this researcher emphasized, "in high-stakes environments, interpretability is just as important as accuracy." Methodically working through seven different cyber attack types, the research sought to combine LLMs with contrastive research (anomaly vs. normal) to produce natural language explanations that help an operator understand quickly *why* a particular flow is suspicious, communicate this to non-technical stakeholders, and improve response time and efficiency.

A US researcher described her organization's work related to AI and cybersecurity adding more layers to the topics mentioned above, such as ways in which adversaries can use AI to augment their cyber-attack tactics. This includes securing the unique system vulnerabilities of AI-enabled systems through such tactics as red teaming to discover vulnerabilities. It is not only about finding out what the technology can do, either on offense or defense, but also how adversaries might choose to utilize these new tools creatively and change the way that attack cells coordinate their activities. Given the fast pace of change in this arena, it is imperative to constantly stay aware of emerging "edge" capabilities, and her organization contributes to this via a lightweight operational evaluation framework that allows cybersecurity experts to contribute to rigorous and repeatable measurement of the plausible risks associated with any given LLM used for offensive cyber operations. This facilitates information and incident sharing and builds a database of knowledge exploitable for LLMs on defense.

An invited private sector participant in one workshop weighed in on the "AI for cybersecurity" discussion and shared his company's experience with AI being used for both offense and defense. He noted that AI is clearly being used effectively by cyber criminals and state-sponsored actors, informing the group that AI-enable phishing attacks are roughly three-times more effective than before. But although defense is inherently more difficult than offense (due to its reactionary nature), AI is bolstering defensive capacity by consuming the vast amount of data available to large firms that could help level the playing field. The largest hyperscalers are processing over 100 trillion signals per day and screening more than 5 billion emails, he said, yielding about 4.5 million malware detections. Compounding daily, this scale of timely data gives AI-enabled cybersecurity systems a valuable edge.

In a recent international AI cyber challenge, for example, the participant described a competition involving the scanning of 54 million lines of code, with the better systems detecting nearly 80 percent of synthetic vulnerabilities (i.e., those introduced for the competition) and even found some unintended vulnerabilities in the software. As AI can facilitate relatively inexpensive fixes for

each vulnerability (in the $100s per patch), AI-enabled cyber defenses have a chance to help keep firms and governments keep pace with would be intruders. It is clear, however, that these benefits of scale are not seamless among different private firms, governments, or between allies, and proactive efforts with creative methods of collaboration might be needed to maximize our collective understanding of evolving threats and our ability to defeat them. Also, as the Claude Mythos example might suggest, a huge database of experience could be less a determining factor in the future if the cyber threat tool is sufficiently capable.

Echoing a call for a greater sense of urgency for AI assurance, a US researcher from a previous NEXT workshop emphasized that "governments are trying to catch up and companies are racing to outcompete each other, and what I feel is being left in the middle are practical measures to ensure that the technology being deployed is safe, reliable, and secure." He added that "we know that these models can be fooled, that we have not solved cybersecurity vulnerabilities, and yet we are rapidly embedding it in commercial and financial applications."

Thus, the AI research community between the US and Japan (and with other nations) needs to help educate policy makers and the public about the technology and associated risks, and collectively we should keep investing time and money in public-supported research settings to advance our ability to manage AI assurance challenges. Our workshop participants were adamant that no country could manage this all on its own, and that partnerships (public-private, but also among friendly nations) would be critical for success.

4. Physical AI Security

Physical AI (i.e., AI systems with integrated mechanical components that perform physical activities with real-world sensor feedback) presents some similar but also many different types of safety and security risks compared to traditional AI systems operating in virtual spaces. Physical AI is gaining attention as private firms expand their integration of AI features with various types of robots and autonomous systems to expand applications. Japanese companies have been leaders in industrial robotics for several decades, but traditionally these were narrowly programmed to perform specific repetitive tasks while incorporating some sensor capacity to maximize performance and safety. However, the rise of AI technology is accelerating hardware-based data collection and utilization, leading to much more capable and complex systems. During our AI Study Trip in Tokyo, we organized one roundtable to consider how each country assesses physical AI risks and what steps they are taking to identify and help manage them, with a goal of sharing information and best practices.

A Japanese participant briefed the roundtable group on recent investment and patent application trends in Japan related to AI and robotics, showing significant growth in recent years with an emphasis on robot systems and control software (e.g., technology for robot motion planning,

trajectory generation, calibration technology, swarm intelligence, and others). These are areas where Japanese firms are trying to improve or gain skills through partnerships, and to some extent they offer in return access to their strengths in key components, such as adaptive and multifunctional gripper technology for diverse objects, among other components. He explained how Japan excels in sensing for safety control, while US and European firms lead in swarm and cooperative control technologies. Along with the robotic systems themselves, companies are investing heavily in data protection and secure operation technology at the robot edge due to cybersecurity concerns.

US and Japanese firms are already working together on different projects that combine their talents including an initiative by Japan's heavy equipment manufacturer Komatsu and Applied Intuition of the US, whose AI and simulation technology will help turn traditional mining equipment into "software-defined vehicles" that can work job sites autonomously. In another example, Japan's factory robot manufacturer Fanuc is collaborating with chip maker Nvidia to develop AI-powered industrial robots that can perform tasks in response to verbal commands. As these types of new inventions come to market over the next several years, policy makers are likely to be confronted with a variety of novel AI-related risks that differ from LLM AI operating exclusively in virtual spaces. In most cases, physical AI introduces a wider array of potential attack surfaces, more potential "weak link" components that need monitoring, and generally higher stakes for danger if systems fail or are compromised (at least in terms of human safety). The testing framework for physical AI risk will need to be more robust than that used for industrial automation control systems, and new standards and regulatory oversight might be required.

A US participant presented some of the research that his team is working on related to future physical AI risks, anticipating continued advancements in AI and robotics technology. Part of their research involves forecasting possible crisis scenarios, such as an adversary gaining control over dangerous robotic systems, and then testing policy makers' responses when faced with different challenges in scenario-based table top exercises. The results highlighted a series of "no regrets" options for action such as hardening certain critical infrastructure, preparing to isolate affected data centers, conducting targeted shut-downs of cyber-physical systems, and strengthening operational dialogue channels between government responders and frontier labs that have the most relevant technical knowledge. He emphasized that safety and security should be designed into new systems from the start, including runtime assurance and other techniques, rather than added later in the process. "Defense can become cheaper than attack," he said, "if we invest now."

Based on a Japanese participant's question about prospects for current AI technology (i.e., LLMs and agentic AI systems) evolving into higher levels of artificial general intelligence (AGI, or AI that can substitute for human cognition on many economically or militarily valuable tasks), the US expert admitted that there is little consensus currently about how and when AGI could emerge, but physical AI systems are likely to be a key channel for AGI's development. This is because many AI researchers believe that AGI will require so-called world modeling to make the advances

necessary for AI to fully understand the physical world it inhabits, beyond its current diet of text, images, numbers, and audio. Physical AI and its accumulation of sensor data over time could help future AI systems experience the world in ways more similar to humans.

Relatedly, a US participant raised the issue of robotic supply chain integrity and reliability as a potential alliance concern, given that a high percentage of components for autonomous systems are made in China, and remote access "back doors" have been discovered in some Chinese robots exported to the United States. A Japanese expert shared this concern and noted his country's efforts to diversify supply chains in order to reduce reliance and risk, but it would take at least five to ten years and possibly longer to accomplish that goal. The US and Japan working together with manufacturers in Southeast Asia could improve chances for success, but there are multiple supply chain initiatives competing for policy makers' attention and appropriators' funds at the moment. This area would benefit from greater focus and strategic alignment.

Shortly after our Tokyo Study Trip concluded, the US Federal Communications Commission (FCC) moved to ban foreign-made drones or critical components for unmanned aircraft systems from sale in the United States (with a limited number of exceptions allowed for certain models). In March 2026 the FCC also banned foreign-made consumer grade wi-fi routers due to the "unacceptable security risk" they pose. There is little to suggest that these moves by the Trump administration were coordinated with allies like Japan in advance (e.g., Japanese systems were initially banned for sale in the US in both cases), and the practical impact of these moves is still being examined. Better inter-governmental coordination of such supply chain steps would help improve preparedness for these kinds of decisions and lessen disruption. Japan's Foreign Ministry created a new economic diplomacy strategy division in 2025 (within the Economic Affairs Bureau) to help coordinate these kinds of policy moves (including policies to promote AI) with other governments and among government, academia, and the private sector.

A US participant noted that both the challenge and promise of recent technological progress is that different technologies increasingly talk to each other and blend together to make new things. "What is autonomy today?" he asked. "It is processing, it is ASICs, low power electronics, localization, 5G communications, it's power and battery technologies, it's navigation, materials, narrow AI, and all sorts of other technologies. And it is happening all over the world under varying degrees of monitoring and supervision." He argued for expanding inter-government connections in the technology space so that they can enhance self-awareness of developments in real time and facilitate communication and coordination during a crisis. His experience in government had demonstrated to him that old style thinking of "throwing people at a crisis" (i.e., deploying thousands of personnel to respond to a natural disaster or a border crisis) could not scale up sufficiently to manage increasingly large and severe events. "Technology can scale more effectively to help meet a challenge, if we can take advantage of it properly," he said.

5. The US-Japan Alliance and AI Assurance

Workshop participants built upon their discussions of current research to explore possible opportunities for collaboration and to share best practices based on their experiences. Bilateral collaboration in AI fields is being stimulated through such channels as the Cross-Pacific AI Initiative (X-PAI) involving US and Japanese universities and companies, a similar initiative involving Carnegie Mellon University and Keio University funded by Arm and SoftBank Group, and Japan's NICT's CREATE program supporting cooperation with US firms and universities, among other means.

NICT in particular is investing time and personnel to support collaboration with US experts, staffing up its office in Washington, DC to become a hub for liaising with US partners to share data and research information, to conduct collaborative projects leveraging shared resources, to organize workshops or joint paper writing, and to host visiting researchers at each other's institutions. NICT has been pursuing similar activities with partners in Taiwan, the United Kingdom, South Korea, Singapore, and elsewhere, all with a common goal of improving their ability to cope with the increasing cyber threats in the domain of AI and cybersecurity.

Because successful collaboration requires partners to understand each other well and develop harmony in their goals and expectations, each collaborative activity offers an opportunity for the partners to connect and improve their communication. Workshop participants noted that overcoming the physical and cultural distance between international institutions is usually the biggest challenge, but these gaps can be bridged with patience, commitment, and institutionalizing these international connections.

A US study trip participant suggested that alliance collaboration include research work on trusted supply chains for both hardware and software relevant to cybersecurity. This could include strategies for ensuring the integrity of data as it is accessed by different systems, and it might also relate to disconnected physical AI systems that are generating data at the edge through sensors during operation. Japan is particularly interested in deploying such systems where domestic labor inputs are lacking (e.g., agriculture, construction, and mining), so it could offer an early proving ground for techniques to utilize edge data safely and facilitate patching for widely distributed systems.

The researcher highlighted some of the collaborations that his US (Department of Energy) laboratory is already conducting with Japanese partners, such as hosting students and visiting researchers from Japan Science and Technology Agency, Japan's Agency for Medical Research and Development, Japan's Space Agency (JAXA), and others. A common challenge emerging from these types of engagements, however, is how to capitalize on their inherent long-term potential value by remaining engaged with this talent after the initial collaboration ends. Developing some kind of alumni network of past researchers to sustain connections or purposefully integrating returning experts with likely future candidates can help build stronger institutional bonds that build over

time. Keeping track of the later achievements of past visiting researchers is also a way to encourage sustained appropriations by demonstrating value for investment.

Participants from both countries emphasized different (sometimes competing) types of collaboration priorities, offering various avenues for research organizations to pursue. Some believed that there was little need to invent possible crisis scenarios or look too far into the future when there were so many near-term concrete problems that need addressing today. These could include new technologies to combat disinformation, better protect vulnerable AI systems, or prevent identity theft or spoofing of identify verification measures. Others argued for at least some focus on longer-range hypothetical problem solving, if only to get a small head start on future challenges to be better prepared. A Japanese participant added that longer-range scenario development can help inform multilateral bodies considering new technology standards. This kind of "peering around the corner" in partnership with industry is the only way that policy makers can contribute productively to future standard setting.

The researchers in our study trip roundtables, and the institutions that they represent, are no strangers to international collaboration. NICT, for example, listed 36 different global partners (in Europe, Asia, North American, etc.) just for its Cybersecurity Research Institute. Most of that collaboration falls into one of four general categories: 1) joint data collection and analysis; 2) researcher exchanges; 3) annual workshops and co-organized events; and 4) joint paper writing. Other examples of US-Japan cooperation include the sharing of national capital-intensive equipment (e.g., lasers or telescopes), joint funding of research by NSF and the Japan Science and Technology Agency, and cooperatively built infrastructure for joint scientific research (e.g., in space). The United States and Japan have been collaborating in various science and technology fields since the early 1960s involving hundreds, if not thousands, of partnerships, and the government-to-government mechanisms for supporting these initiatives are starting to grow.[5]

Still, although the basic challenges and opportunities related to international collaboration are well understood, technical collaboration related to AI assurance is in its early stages, and even if larger parent organizations in both countries are familiar with each other, the researchers and their labs that focus on AI issues might not be aware of those connections and best practices. Participants from both nations with significant experience in collaborative international science projects offered their insights to the group.

There are a variety of options for bilateral collaboration. One approach is to divide up a research task into two separate activities that play to each other's strengths, and then combine the result from these different capabilities. Another option could be having one partner "red team" the results of the other partner, as is conducting similar research and then comparing the outcomes.

[5] The last two pages of this report draw heavily from our previous Study Trip summary in 2025 in the section on "Best Practices for International Collaboration" (pp. 19-21). It is updated with some additional insights from our December 2025 study trip.

However it is handled, an additional expected benefit from such efforts should be an improved ability to advise policy makers in both countries to make difficult alliance decisions that need scientific input. This was the case, for example, during the COVID pandemic or when assessing potential health risks and evacuation zones during the Fukushima nuclear power plant disaster in 2011 in Japan and bilateral activities were carried out for disaster relief and nuclear risk management.

At the workshops, participants identified a variety of research areas that could be good candidates for bilateral collaboration including deepfake detection technologies, originator profile technology, cybersecurity protection and malware detection technologies (understanding co-generation principles), software supply chain reliability and resiliency, federated cloud technologies (and the balance between cloud computing and edge computing), and the area of human-digital twins, which could corelate with some of Japan's work on personal AI technologies.

Overall, participants advised a "defense in depth" approach for AI assurance, recognizing that a gap or problem could arise at some level within the system, but with sufficient depth it could be stopped at the next level or a level beyond that. One researcher likened this to Swiss cheese that might have holes at any given location but not all the way through. Still, a solid foundational layer for the technology (and its underpinning data and performance standards) is vital as AI scales out more broadly into a wider range of uses. Building consensus internationally around this foundational layer (and how to communicate when problems arise) struck some participants as an important area of collaboration. Part of this could include coordinating on data governance frameworks for AI. Mitigation techniques, promoting AI literacy, and workforce development were not discussed in detail during the trip, but they could be other good areas for bilateral collaboration.

As mentioned in our previous report, in addition to bilateral collaboration at the individual or institutional level, a pair of government-convened committees is an important overseer of US-Japan science and technology cooperation, and they have placed AI high on the list of policy and research priorities. The Joint High-Level Committee (JHLC) and the Joint Working-Level Committee (JWLC) on Science and Technology Cooperation were formed under the 1988 US-Japan Agreement on Cooperation in Research and Development in Science and Technology. The JHLC (and the JWLC that supports it) have a complicated history and a complex governing structure, which have contributed to its lack of ability in the past to be a strategic control tower for allied science investments, but that could be changing. Also, even if the JHLC has a limited ability to organize and manage bilateral science and technology cooperation, it does influence priorities and provides a bird's-eye view of all allied activity from which all stakeholders can benefit.

On the US side, the JHLC is chaired by the OSTP Director but includes active participation from a variety of interested offices and departments including the Department of State, the National Institute of Standards and Technology (NIST), the National Oceanic and Atmospheric Administration (NOAA), the National Science Foundation (NSF), and the Department of Energy

(DOE). The Japan side is co-chaired by the Minister of State for Science and Technology Policy, Intellectual Property Strategy, Space Policy and Economic Security, and the Minister of Education, Culture, Sports, Science and Technology (MEXT). Other participating Japanese institutions include the Cabinet Office (CSTI), Ministry of Internal Affairs and Communications (MIC), Ministry of Foreign Affairs (MOFA), and Ministry of Economy, Trade and Industry (METI).

The two committees usually meet annually or slightly less often, though during the Biden administration they were becoming a more valuable means for inter-agency coordination of bilateral science and technology initiatives, especially as they relate to AI, quantum computing, space, climate change and health. The return of the Trump administration, however, has changed the dynamic and apparently pushed the JHLC to the side as a means of coordinating bilateral basic science collaboration. Instead, OSTP has played a central role and technology cooperation has become linked with Trump's broader trade and economic security agenda.

This was evident in the "Technology Prosperity Deal" announced when Trump visited Japan in October 2025, a document that announces a wide range of science and technology priorities for the allies to pursue as they see fit, with implementation (and funding) left up to individual agencies. The foundational 1988 US-Japan Agreement on Science and Technology was in danger of expiring in early 2026, but it was extended for nine months from January 20. This kind of short-term extension is not conducive to long-term planning for science and technology collaboration.

With a technology as impactful and fast-evolving as AI, every year is a kind of crossroads for the alliance in terms technology and policy coordination. US and Japanese policy makers and research organizations should be accelerating their engagements related to AI and its potential effects on our economies and national security, but US-Japan institutional and human networks in this field are still underdeveloped. Although it is difficult for the two countries to significantly expand investment in bilateral research collaboration due to budget limitations, they can make better use of activity already underway through greater communication, networking, and focused attention on building relationships that deliver mutual benefit. Existing frameworks should be leveraged for maximum contact and output, which in turn should signal where the most promising returns can be found in the future. Our "Allied on AI Assurance" project will continue to foster networking and information sharing on this front throughout 2026 and 2027.

NB: Please note that this report is not a consensus document and is solely the author's own interpretation of the insights generated by the dialogues and meetings in 2024 and 2025 (and via subsequent research) regarding AI assurance and the US-Japan alliance. Quotes are used to illustrate points on a not-for-attribution basis, as per agreement with all the dialogue participants. The author is solely responsible for the content of this report. Questions or comments regarding this summary can be directed to Jim Schoff at jschoff@spfusa.org

About the Author

Jim Schoff is currently a distinguished senior fellow (non-resident) at Sasakawa Peace Foundation USA. His career spans 38 years working in the fields of business, education, government, and the non-profit sector, all related to Japan, East Asia, and the US-Japan alliance. From 2021-March 2026 he was senior director at Sasakawa USA and led the foundation's "US-Japan NEXT Alliance Initiative," which mixes bilateral dialogue and research to stimulate new alliance connections across foreign, security, and technology policy areas. Schoff is also president of the consulting firm Asia Defense & Alliance Management (ADAM), Llc and Chairman of the Board of Trustees, Japan-America Society of Washington, DC. At Sasakawa USA, Schoff continues his work on artificial intelligence and the US-Japan alliance with a quarterly bilateral roundtable and related research and interviews on AI governance, safety and security. Previously, Schoff was a senior fellow and director of the Japan Program at the Carnegie Endowment for International Peace for nine years, following two years as senior adviser for East Asia policy at the U.S. Office of the Secretary of Defense and seven years as Director of Asia-Pacific Studies at the Institute for Foreign Policy Analysis. Some of his publications include: "Updating Alliance Strategies to Sustain Nuclear Deterrence" (Sasakawa Peace Foundation USA, April 2025), "Allied on AI Assurance: Technology to Enhance AI Security and Safety" (Sasakawa Peace Foundation USA, October 2024), "Modernizing US-Japan Command & Control Relationships for New Challenges" (Sasakawa Peace Foundation USA, May 2023), "US-Japan Technology Policy Coordination: Balancing Technonationalism with a Globalized World" (Carnegie, 2020), and *Uncommon Alliance for the Common Good: The United States and Japan after the Cold War* (Carnegie, 2017).

---

*The **US-Japan NEXT Alliance Initiative** is a forum for bilateral dialogue, networking, and the development of joint recommendations involving a wide range of policy and technical specialists (in and out of government) to stimulate new alliance connections across foreign, security, and technology policy areas. Established by Sasakawa Peace Foundation USA with support from the Nippon Foundation, the goal is to help improve the alliance and how it serves shared interests, preparing it for emerging challenges within an increasingly complex and dynamic geostrategic environment. Launched in 2021, the Initiative includes two overlapping lines of effort: 1) Foreign & Security Policy, and 2) Technology & Innovation Connections.*