Medium-Term AI Risk Foresight

March 31, 2026

(Revision 1.0.0)

Committee for Machine Learning Quality Management

Artificial Intelligence Research Center

Technical Report

Cyber-Physical Security Research Institute

Technical Report CPSEC-TR-2026002

Intelligent Platform Research Institute

Technical Report IPRI-TR-2026-02

National Institute of Advanced Industrial Science and Technology (AIST)

Executive Summary

The rapid acceleration of AI development is outstripping the ability of societies to update laws, standards, and governance frameworks. While traditional IT matured slowly enough for regulation to keep pace, modern AI—particularly generative, agentic, and potentially autonomous systems—evolves too quickly for reactive, incident-driven approaches. This report examines how AI-related risks may emerge and propagate into harm, and on that basis sets out a medium-term research agenda identifying the areas in which further work should be prioritized.

A central feature of the report is the AI Risk Materialization Framework, which identifies five interacting factors: development and operation, societal governance, AI use, the relationship between AI and the world, and the resulting harms. Harm arises only when multiple factors interact, underscoring that AI risk cannot be reduced to technical defects alone but must be understood as the product of systemic conditions surrounding AI systems.

The document analyzes how these factors manifest across major AI categories. Predictive AI presents comparatively narrow risks due to its constrained outputs and expert user base. Generative AI introduces richer pathways to harm: hallucinated factual content, creation of misleading media, exploitation by malicious users, psychological influence in companionship settings, and technical vulnerabilities in engineering design tasks. These risks vary significantly with user expertise, dissemination channels, and the contexts in which outputs are used.

Agentic AI represents a sharper escalation. By autonomously operating external tools, interacting with physical or cyber systems, and executing transactions, agentic systems can produce real-world effects without sustained human supervision. Their behavior in open environments—with unclear authorization boundaries and unpredictable interactions with external actors—creates qualitatively new governance challenges. The report also anticipates AI saturation scenarios in which large numbers of AI agents interact in unpredictable ways, generating system-level overload or unintended coupling across infrastructures. Looking further ahead, the document considers the emergence of autonomous entities pursuing self-defined goals and stresses the need for research to prevent harmful autonomy while understanding how intentionally built systems might be made safe.

A review of existing international taxonomies shows that current lists of “AI risks” often mix harms, system behaviors, and usage issues without distinguishing their causal roles. Mapping

these categories onto the proposed framework clarifies where current classifications fail to capture mechanisms that actually lead to harm, limiting their utility for governance or standard-setting.

Considered through the AI Risk Materialization Framework, the evolving scenario space of AI risk highlights a set of research challenges that are likely to become increasingly important as AI systems expand their range of action, interact with more varied environments, and continue to evolve technically.

The research agenda is structured along four axes. First, present AI systems struggle with unstated assumptions, limited understanding of real-world regularities, and insufficient generalization across domains and unfamiliar situations. Second, AI must be capable of handling cultural, environmental, or societal contexts rather than assuming uniformity across regions. Third, safety responses must integrate technical safeguards in line with broader organizational procedures and societal arrangements, given that many risks arise from interactions beyond the developer's control. Fourth, safety approaches must be resilient to discontinuous changes in AI architecture, ensuring that core principles remain valid even when technical foundations change.

In sum, the report calls for proactive safety research that anticipates risk pathways before they solidify, integrates technical, organizational, and contextual considerations, and develops principles that remain robust amid rapid and potentially discontinuous technological evolution.

k.konishi@aist.go.jp
デジタル署名者 : k.konishi@aist.go.jp
日付 : 2026.03.31 15:40:36 +09'00'

1 An Overview

The process of creating rules such as laws, guidelines, and international standards for engineering products, including those involving artificial intelligence, generally takes many years. This was acceptable for conventional IT systems and software; these technologies took decades to reach maturity, allowing plenty of time for the creation of rules to catch up with their development. Unfortunately, this is not the case with artificial intelligence, the development of which is accelerating to an unprecedented level. Such rapid development leaves little time for the development of rules if they are started after incidents caused by this technology occur, or even after the technology becomes available. In this document we attempt to predict the emergence of new kinds of risk that might be caused by AI evolution in coming years and explain our intention to focus on some of these risks in our medium-term activity.

2 AI Risk Materialization Framework

In this document, we propose a plausible assumption concerning how AI can be involved in the materialization of risks into harm. This assumption is illustrated in Figure 1. This assumption should guide our understanding of the different types of risks and their characteristics. Such understanding should then guide us in identifying measures to prevent those risks from materializing. It is important to note that some combination of factors may not necessarily lead to any risks or harm even when instances of the factors in such combinations are potentially dangerous.

The following five types of factors for risks of AI are distinguished:

1) factors during the AI development and operation,
2) factors in effort for societal governance
3) factors during the AI use,
4) the relationship of AI and the world, and
5) harms.

The combination of these factors will lead to a specific type of risk scenario. Factors 1-3 are found in and around AI systems. Factors 4 and 5 manifest in the world, in which AI systems operate. We acknowledge that the conventional hazard theories [1][2] typically identify three components in hazard manifestation: source, trigger, and outcome. We respect this conventional approach. We also note that scenarios in which AI systems cause harm may have multiple

branches and each branch can be lengthy. In other words, a scenario leading to an outcome may involve multiple sources and triggers. To focus on typical scenarios while maintaining sufficient details, we distinguish two domains in AI risk manifestation scenarios. The first domain concerns AI systems and contains sources and triggers that lead to a problematic event with AI. The second domain encompasses AI and the world and the triggers in it entail harm to actors in the world. Whether or not triggers in the first domain lead to harm depends on triggers in the second domain.

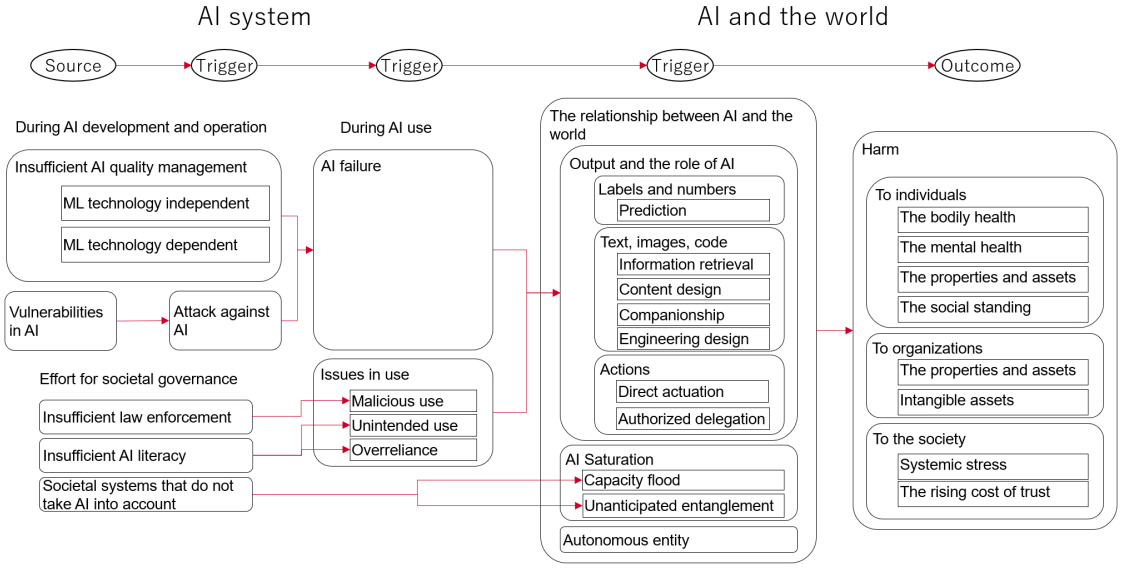


Figure 1 AI Risk Materialization Framework

2.1 Factors during development and operation

The factors during the development and operation of AI systems can be divided into two; (1) insufficient quality management performed by the developer or provider and (2) vulnerabilities in AI and attacks exploiting them against AI.

Insufficient quality management refers to any deviation from the right process and states of the items involved during the development and operation of AI, and it typically happens because of the lack of attentiveness by the developer or provider of the AI. Those items involved and the deviations associated with them are either technology-independent or technology-specific. Technology-independent items are activities that are necessary regardless of the type of AI technology used to build the system, whether it be rule-based, machine learning, generative models, or what may come next. Examples are requirement analysis and risk analysis. Technology-specific items are activities that are necessary only to build the

system based on specific types of AI technology. Dataset design and management are necessary only in building systems based on machine learning and later technologies. Prompt engineering is necessary only for generative AI and later technologies.

Vulnerabilities in AI make it possible for attacks against AI to bring about deviations from the right procedures and states involved in the development and operation of AI. Such attacks are conducted by malicious actors in most cases. Such actors may be operating externally or internally to the developer or provider of the AI. Malicious actions causing defects include poisoning of various data such as datasets for training and testing and system prompts for generative AI, as well as any hindrance to correct model training. In addition, indirect prompt injection may occur during AI operation. For example, if AI incorporates rigged search results into its prompt when searching the web, it may follow the instructions of whoever rigged the web content.

2.2 Factors in effort for societal governance

The factors in the effort for societal governance over AI concern the situation surrounding AI. One example is society's readiness to combat the malicious use of AI. Such use becomes more likely when there are insufficient regulations and law enforcement. Another example involves preparedness of AI users. Insufficient AI literacy can result in inappropriate use of AI, which neither the designer nor the developer intended. A third example is societal system's readiness for AI. Most societal systems are designed for humans and do not take AI into account. When humans empowered by AI or AI agents participate in these systems, they may be overwhelmed by the flood of AI-generated output and actions and rendered useless. In addition to such super-human productivity, AI-assisted humans or AI agents may exhibit super-human versatility and multitasking capability. When they operate on independent systems simultaneously, they may inadvertently make new interdependence across those systems and cause unanticipated entanglement.

2.3 Factors during AI use

The factors during AI use include AI failures and issues in AI use.

2.3.1 AI failures

AI failures may be the output or behavior of AI that does not align with the requirements as perceived by the developer or provider of AI. For example, output of AI that is incorrect or unsuitable to the situation that the AI must deal with, especially output that is in violation of safety requirements for the AI. Biased or toxic output and output that infringes privacy or copyright can also fall into this category. Note that such output can occur without the user requesting it.

2.3.2 Issues in use

Issues in AI use include inappropriate use of AI, with or without malicious intention. Malicious use of AI, such as use with deliberate intent to commit privacy or copyright infringement, is of course problematic, but even without any malice, use of AI outside the scope of the intention of the developer or provider of the AI may have unexpected consequences. Besides, repeated use by the same individual may lead to overreliance.

2.4 The relationship between AI and the world

The relationship between AI and the world concerns the way AI interacts with the users and other actors that can be influenced by the AI directly or indirectly. The most common form of interaction is for AI to play a role that is useful for the user. Regardless of its specific roles, the imminent proliferation of AI is likely to change the way societies work. Furthermore, AI that acts independently of human instruction — autonomous AI — may emerge within this decade. The relationship between AI and the world determines whether any harm results from problems during AI use, and what kinds of harm may result.

The roles of the AI refer to the types of functions the AI should exhibit in its use. Along with the development of new AI technology, the roles of AI have expanded and will continue to do so. It is important to note that we should also consider the role of AI played in its abuse by malicious or inadvertent users. Although the conceivable roles of AI are quite diverse, they can be chunked into categories. The list below attempts to categorize such roles. While it is not exhaustive, we hope that it covers most conceivable roles.

- **Prediction** — provide predictions based on input and pre-determined and/or limited

criteria

- **Information retrieval** — provide and pre-process information to assist humans' acts.
  - Summarize, translate, or transform given input.
  - Gather information from external sources.
- **Content design** — produce or help produce artistic or communicative materials such as literature, art, music, manuals, announcements and reports.
- **Companionship** — influence humans via visible, audible or subliminal means such as counselling, therapy, or fortune-telling.
- **Engineering design** — make parts used in the industrial systems. It can perform tasks such as programming and 3D printing tasks, for example.
  - Build computer software.
  - Build physical objects (for example, parts of a mechanical device for replacement)
- **Direct actuation** — operate machinery or information systems to perform actions that result in a change of state.
  - Act in a closed space
  - Act in open spaces
- **Authorized delegation** — make contracts and promises, pay or receive money, negotiate with others (humans or AI agents) on the user's behalf.
  - Negotiate with (conventional, non-AI) IT systems.
  - Negotiate with humans.
  - Negotiate with other AI delegates.

The roles in the list above are roughly in chronological order. Before generative AI, the only role performed by AI was prediction, where the AI output was either labels and/or numbers. However, generative AI has made it common for AI to be used for information retrieval, companionship, content design, and engineering design, where AI is producing text, images, and code. As of 2025, both academia and industry are keen to utilize AI for direct actuation and authorized delegation, with limited success so far, but rapid progress is ongoing. Here, AI produces actions as its output.

As AI becomes more versatile, it will proliferate and permeate every aspect of society. The large-scale deployment of AI will inevitably change multiple aspects of society, sometimes for the better and sometimes for the worse. For example, AI may be used in quantities that were unforeseen by the developer or provider and disrupt societal systems. Even when a single use of AI is quite appropriate, or at least acceptable, massively concurrent use of the same AI may cause problems by pouring a large amount of output to systems that are not designed to process such a large load. In another case, a single AI agent may operate multiple independent systems and inadvertently and effectively connect them, for example, by feeding output from one system

to another, or regulating one system based on output of another. Such linkage may also happen due to human operators working on multiple systems, but the links realized by AI can be much more intense than those created by human operators.

Eventually, AI may operate autonomously, detached from any human-designated goals. This is still a futuristic possibility but a plausible one. Someone with sufficient resources could build infrastructure to sustainably run such an AI, purely out of curiosity. Alternatively, an AI pursuing human-set goals could be derailed by accidents or cyberattacks and begin conducting self-appointed missions.

2.5 Harms

Harms may be suffered by individuals, organizations, and societies. They can be physical, psychological, economic, or more abstract, for example, concerning reputation and trust.

3 Risks in literature

Guidelines and reports on the management of AI quality, safety and risks are issued by various organizations around the world, including government bodies, academic institutions and private sector companies. These usually provide a list of AI-related risks [3][4][5][6][7]. However, these lists often include items of quite different natures. Some describe harm suffered by people, some describe AI behaviors that do not meet the expectations of their developers or providers, and some describe problematic AI usage. A few examples will be given later in this section. To prevent harm, it is important to recognize the nature of these so-called 'risks' and identify their relationship to the scenarios and situations that lead to harm. This is because measures to prevent harm depend on the nature of the risks.

The MIT AI Risk Repository [3] is a substantial online database of AI risks. A report and an accompanying slide set describing the repository are also available on the website. The report provides AI risk taxonomies. These come in two flavors: causal taxonomy and domain taxonomy. The causal taxonomy categorizes risks based on the entity, intent, and timing associated with each risk. The domain taxonomy, on the other hand, groups risks into seven domains representing different types of potential problems.

The three aspects of risk in the causal taxonomy are certainly important. Our AI Risk Materialization Framework also makes distinctions based on these aspects. Distinctions based

on entity can be found in the following parts in the framework. AI failure represents the cases where risks are caused by AI, while Issues in use indicate the cases where risks are caused by the users. The distinction based on intent appears in the following part. Attack against AI and Malicious use represent intentional risks, while the others represent the cases where risks are unintentional. Finally, the distinction based on timing can be observed in the following parts. Insufficient AI quality management and Attack against AI correspond to situations that occur before AI deployment. The others occur after the deployment.

The domains listed in the domain taxonomy are intended to represent different categories of risk. However, they represent various types of risk factor of different natures. Some relate to harm suffered by something, while others represent problems with AI that may or may not result in harm. Some relate to the use and misuse of functioning AI systems.

Figure 2 illustrates how the risk domains from the domain taxonomy are mapped onto AI Risk Materialization Framework. As can be seen on the far right of the figure, some of the domains represent different types of harm caused by AI. Others, however, concern AI failures, insufficient AI quality management, or issues in use. A few do not map onto the framework at all because they relate to any risk, if at all, only weakly.

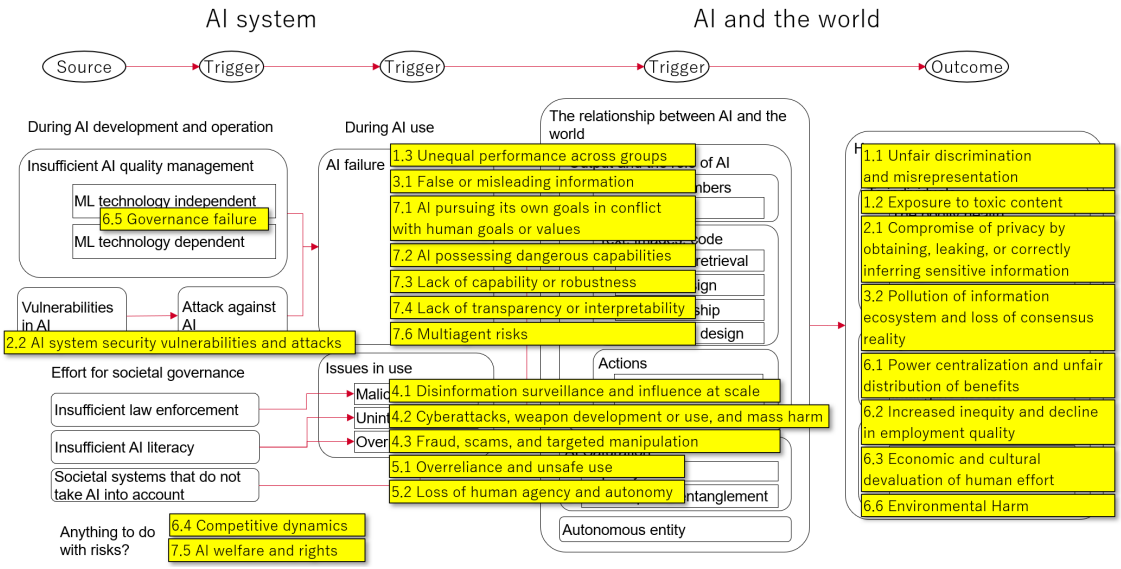


Figure 2 Mapping between the risk domain and AI Risk Materialization Framework

4 AI Medium-Term Risk Prospect

In the AI Risk Materialization Framework, we separated Harm from AI failure and Issues in use and placed the Relationship between AI and the world in between. We did so because we wanted to show that whether any Harm arises from AI failure or Issues in use depends on the Relationship between AI and the world. In the rest of this section, we will analyze what kind of relationship translates AI failure or Issues in use into Harm, and what does not.

4.1 Predictive AI (ETA: 2010s)

Predictive AI based on machine learning only produces labels or simple numbers. AI failure in predictive AI is very likely to cause harm. The output must be correct, robust, and unbiased; otherwise, it can be harmful. However, it is rather unlikely that issues in the use of predictive AI will cause harm. There are a few reasons for this. First, its output is too simple to express complicated information that can result in privacy violations or copyright infringement. Second, predictive AI is usually used by professionals, who are unlikely to put it to unintended use or fall in the state of overreliance. Third, it is rather hard to use predictive AI with malicious intent. Its output is dull and unlikely to cause any intense reactions from the recipient unless presented by the associated experts. Therefore, any malicious use of predictive AI by non-experts is unlikely to cause significant harm. Conversely, an expert using predictive AI for malicious purposes would attract unwanted attention, which would deter such behavior.

4.2 Generative AI (ETA: early 2020s)

Generative AI produces much richer, expressive output such as text, images, and code than predictive AI. On one hand, the high expressivity of its output unlocks unwanted possibilities for malicious and unintended uses. Generative AI may produce private or copyrighted information if such information was used in training the generative AI model. It can be used to produce fake images that look so real and are hard to distinguish from real images. On the other hand, just producing such information will not necessarily cause harm. Such information causes harm only when disseminated widely. In this respect, private information may be an exception. It can be problematic even when just produced and seen only by someone who is not supposed

to see it. Still, just producing such information does not immediately cause harm. Furthermore, the output of generative AI can be so attractive that it can cause addiction and overreliance.

With such expressive output, generative AI can play so many different roles. Here we focus on three of its major role categories: information retrieval, content design, and companionship.

4.2.1 Information retrieval

Generative AI can sometimes be used for information retrieval. It certainly knows a lot of facts, and answers questions correctly when it happens to know the right answer. It is also notorious for its hallucinations, that is, giving wrong answers in disproportionately confident-looking text. That counts as a type of AI failure. However, whether hallucinations cause harm depends on several factors. First, no harm is likely to arise when the user is an expert on the domain of the question. As an expert, she or he can assess the likelihood of the correctness of the answer and, if it looks dubious, she or he can double-check it. In other words, it would be a dangerous practice when the user knows little on the domain of the question, for example, for children to learn a new subject. Second, no harm will arise when the AI is used for creative purposes, which leads us to the next topic.

4.2.2 Content design

Generative AI can also be used for content design. Sometimes it is used to write a formal report that requires factual information. Sometimes it is used to write a fictional story. In either case, an AI failure, such as hallucination, is unlikely to cause harm because users must know what information they want included in their content. For factual documents, users are responsible for ensuring that the AI output is truthful. If the AI hallucinates, the user must recognize it and correct it. Note that if users do not know whether the AI output is truthful, they are essentially using the AI for information retrieval, not content design. For fictional documents, however, there is no obligation to ensure the veracity of the content. The important thing is whether the AI-produced content aligns with what the user wants to express.

While Generative AI used for content design does not easily cause harm when it exhibits AI failure, it easily causes harm when there are issues in use of it. For example, malicious users are using generative AI to produce fake content and then disseminate it to harm people or

organizations by smearing their reputations. This can also lead to societal disruption, for example, by influencing election results.

4.2.3 Companionship

Generative AI can also be used to provide companionship. For example, there are merchandised products that are called “social robots” or “robotic companions”, which interact with the user typically in friendly fashion, often in voice. Not just AI in the form of robots but so called “chatbots”, including ChatGPT, are often used for companionship rather than for business or educational purposes. For companionship, correctness of the AI output is not the most important quality. In some cases, “gentle lies” may be what the user needs most at that moment. In contrast, friendly, heart-warming, cheerful, and caring interactions are some of the most expected features of AI used for companionship. Deviations from those features count as AI failures. For example, aggressive, toxic, depressing, or cold interactions. Whether such AI failures lead to harm depends on users’ mental toughness. A sufficiently mature and healthy mind would just get angry at AI behaving like that and throw it away. However, sensitive minds such as teenagers’ or mentally stressed can be shocked and damaged by AI’s such behavior.

Additionally, well-functioning AI designed for companionship may cause overreliance, a type of issues in use. Users may become addicted to AI and be unable to stop using it. Users may also be encouraged to take actions they would not otherwise take, which can result in irreversible tragedy. Whether or not overreliance leads to such harm depends on the mental maturity and conditions of the users.

4.2.4 Engineering design

Another important application of generative AI is engineering design, including program coding. Programming with generative AI is already popular and is known as "vibe coding." Another example is using generative AI to design 3D parts for 3D printing. These practices are much easier for people with no prior experience in programming or part design. More and more people are doing so, with some doing so habitually and others only occasionally. This is yet another example of the democratization of professional skills brought about by AI, which will greatly promote innovation in society at large. However, these practices are highly vulnerable to AI failures caused by insufficient AI quality management or attacks against AI.

Imagine that a bolt on a large shelf in your child's room breaks. You could take a picture of the broken bolt, upload it to a chatbot, and have the chatbot design a replacement. Then, you print the design with a 3D printer. The new bolt fits the shelf perfectly, and the shelf is fixed. Or is it? Neither you nor the chatbot knows precisely how much load-bearing capacity the bolt needs to possess. Consequently, you may have 3D printed it with the wrong material such as a type of plastic which is hard but brittle.

4.3 Agentic AI (ETA: Mid-to-late 2020s)

The definition of AI agents and agentic AI is still a topic of discussion. In this document, the term is used to describe AI with capabilities to take actions, that is, to operate external tools to obtain information, typically live, dynamically changing information, and/or to make changes to the state of external objects or systems, both physical and cyber. Such capabilities unlock a wide variety of uses, including manipulating vehicles, robots, and other machinery (direct actuation), as well as booking and purchasing tickets, hotels, and other merchandise on behalf of the user (authorized delegation). Since these capabilities directly affect the state of the world, AI failure or issues in use of agentic AI can lead to harm more easily than those of predictive and generative AI.

4.3.1 Direct actuation

AI agents perform actions in the world without relying on human assistance or intervention. These actions may cause harm, which can be induced both by AI failure and by issues in use. AI agents have the potential to cause more harm than predictive and generative AI because they are designed to work without human supervision. To ensure safety, it would be helpful to introduce human oversight of AI agents. For example, humans could oversee every action an AI agent takes and judge its safety. Doing so, however, would defeat the primary purpose of deploying AI agents, which is to free humans from performing those actions and to liberate the system from the speed limit of human reflexes.

4.3.2 Authorized delegation

The severity of the harm caused by AI agents depends on the scope of the operations they perform. The potential for harm is limited when an AI agent operates in a sandbox—a contained

environment outside of which whatever happens inside has little influence. Operations in a sandbox are safe because everything in it is under control of a single party, which may be just you or a project team across organizations. There can be a prior agreement within that party that anything in the sandbox may suffer damage. Such an agreement would make the situation much easier to handle when something goes wrong.

In contrast, the harm caused by AI agents can be much trickier when they operate in an open world consisting of assets owned or controlled by external parties. Complication can arise because that implicitly and effectively involves striking a contract without your consent if the AI agent did so due to AI failure.

For example, you may want your AI agent to book a hotel room for your next business trip. This room is the property of the hotel, not yours. There is no prior agreement between you and the hotel because your AI agent found the room via a web search. Nevertheless, by booking the room, the AI agent makes you responsible for paying for it. If you like the room, that's great. Otherwise, you will be in trouble. The risk is relatively small in this case, as hotel room bookings can often be canceled without a cancellation fee until a few days before the stay. The cancellation policy may be less flexible when purchasing other goods and services.

Current e-commerce practices assume that you, the user, are a human and not an AI agent. Although they verify your identity and payment capability, they do not check whether you have authorized someone else, an authorized delegate, to act on your behalf. They also do not verify the scope of the authorization, including permitted actions and credit limits. This assumption will change in a few years as AI agents will become more widely deployed. This new assumption will require new protocols to securely demonstrate and verify the scope of the authorization.

4.4 AI Saturation (ETA: late 2020s)

AI will saturate the world, and it will be filled with a lot of unforeseen situations. As AI becomes more widely deployed, it will appear in scenes and situations where it is not expected. This will create mismatches between conventional assumptions about human behavior in these situations and the actual AI behavior. In addition, as more AI agents operate in the same space—whether physical or cyber—they will encounter each other more often by coincidence. No one will anticipate, plan for, or design protocols for these encounters.

A type of mismatch will be quantitative. AI can perform actions thousands or even millions of times faster than humans, which creates a capacity flood. Although that has been the case also with more conventional IT systems, what is new is that they will be able to do so while interacting with the systems via UI designed for humans. Besides, although DDoS attacks against web sites have been around for years, the new threats enabled by AI will consist of requests that are, when examined individually, completely legitimate, issued by independent AI, coordinated by nobody. Such requests can be numerous because there are many AI agents, and each of them will be able to issue much more requests faster than humans conventionally do.

Another mismatch will be qualitative. Some AI, such as multimodal generative models, can handle multiple modalities simultaneously and rapidly identify correspondences between features in one modality and features in another. These models may accidentally create unanticipated entanglement between the states of independent systems, observed in separate modalities. This entanglement may introduce unwanted causality between those systems.

It should be noted that harm may result from well-functioning AI when it saturates the world. This can occur when there are no AI failures, attacks against AI, or issues in use. Key sources of this risk scenario are societal systems that do not take AI into account.

4.5 Autonomous entity (ETA: 2030s or later)

The word “autonomous” has different meanings. When people talk about an autonomous vehicle, they mean it as a tool that assists humans. It receives instructions from humans about where to go and then it goes there. It is considered autonomous because it does not need to be told how to get there. However, it still follows human instructions. In contrast, we can imagine an entity that follows no human instructions and instead pursues its own goals. This entity is no longer a tool. This section discusses entities that are autonomous in the latter sense.

The idea of AI being autonomous and not following human instructions sounds scary. However, people are autonomous, and yet they follow certain rules, form societies, and participate in them. They do so in some cases because of coercion, but in many cases spontaneously. The question is whether autonomous AI pursuing its own goals can work with humans to create a harmonious society. To answer that question, we have a lot of work to do.

Autonomous AI can arise by accident or by design. By definition, accidental autonomous AI does not guarantee that it will function as its developer intends. Therefore, it is very unlikely to

work harmoniously with humans. What needs to be studied about it is how to prevent its occurrence. Autonomous AI by design can be divided into two categories: disruptive and harmonious. Disruptive autonomous AI may result from cyberattacks targeting tool AI that pursues human-set goals. Such AI is most likely intended to cause harm, so it does not make sense to consider how to make it safe. Harmonious autonomous AI would result, if at all, from someone’s desire to develop such an entity. The developer's motivations are anyone’s guess: technical curiosity, the belief that it will help humanity, a lust for fame as its inventor, or something else entirely. At least one thing is clear. By definition, the developer wants it to be harmonious. This justifies and warrants research to make that possible.

5 Our Medium-term Research Agenda

AI safety research must address not only the limitations that can already be recognized in present-day systems but also the possibility that the assumptions behind current safety work may change as AI continues to evolve.

The first section concerns limitations that cannot be fully addressed by collecting more examples or extending lists of known scenarios. The second concerns additional difficulty arising across regions, cultures, and social contexts. The third section considers how such issues must be addressed across different layers of response rather than within technical measures alone. The fourth section then turns to the need to prepare for changes in the technical foundations on which current safety work depends.

The research agenda catalog that follows is organized on the basis of this structure.

The relation between this research agenda and the AI Risk Materialization Framework requires clarification. The framework is intended to enumerate broad classes of risk materialization scenarios by showing how harm may arise from combinations of factors in development and operation, societal governance, AI use, and the relationship between AI and the world. The research agenda is not organized as a direct mapping onto those elements. Instead, it identifies structural bottlenecks that are likely to become more important as AI systems expand their range of action, interact with more varied environments, and continue to evolve technically. This relation is illustrated schematically in Figure 3. The four groups of research issues in this chapter should therefore be read as cross-cutting priorities derived from that scenario space, rather than as a one-to-one restatement of the framework.

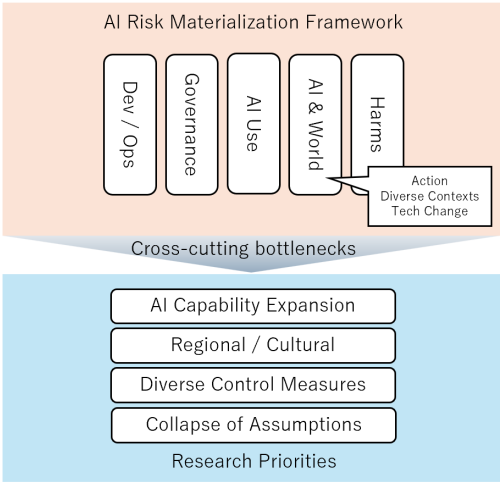


Figure 3. From the AI Risk Materialization Framework to Research Priorities

5.1 Safety Limitations Under Expanding Capability of AI Actions

Many of the safety problems discussed in this section reflect limitations that become more serious as AI systems expand their range of action and operate across more varied and less predictable environments. Present AI systems learn mainly from explicit data and documented cases, but safe behavior in the real world depends on conditions that are often left unstated, cannot be exhaustively enumerated, or cease to hold once AI operates beyond a single application domain.

A first difficulty lies in hidden real-world assumptions that people rarely bother to write down. Everyday life depends on constraints that are so obvious to humans that they are often not expressed as rules at all. A simple example is a physics law that an egg,a glass, or a soap bubble will break if it is gripped too strongly. Another is legal laws that even when a rule appears to be written down, such as a traffic rule or a legal requirement, its correct interpretation often depends on background understanding that is not fully stated in the text itself. Present AI systems mainly learn from material that is explicitly given to them in documents or training data, which makes such unstated conditions difficult to acquire in a reliable way.

A second difficulty is that real-world safety cannot always be reduced to a collection of cases. In some restricted applications, large numbers of simulations and accident examples can still support practical progress. Autonomous driving is a good example: at present, it may be more practical to learn avoidance behavior from driving simulations and accident footage than to build a system around an explicit, principled understanding of every physical and human factor

involved in safe driving. Yet this approach becomes much less sufficient in settings such as robots acting in everyday human environments, where it is not realistic to prepare training scenarios for every event that may occur. In such settings, the issue is no longer whether the system has seen enough examples, but whether it has enough understanding of how the world itself behaves to act safely when no close precedent exists.

A third difficulty is that safety knowledge organized by application area becomes insufficient as AI systems act across a wider range of situations. Current practice often teaches hazards separately for each domain, such as driving, service robotics, or software assistance, and this remains workable while the scope of action stays narrow. The problem arises when those boundaries no longer hold. A robot designed and trained mainly for domestic assistance, for example, may eventually encounter situations that require it to interact with transportation systems, public infrastructure, or other environments that were treated as separate domains during development. In such a case, safety knowledge prepared only for household robotics, or only for driving, would not by itself be enough. The difficulty is not simply that one more scenario must be added, but that the system has entered a situation where the boundaries between previously separate domains have become unstable.

Taken together, these issues point to a broader limitation of present AI safety work. Some necessary conditions for safe behavior are not written down, some cannot be covered by examples alone, and some emerge only when the system's range of action expands beyond a single field. For this reason, the safety of future AI cannot be treated simply as a matter of adding more data, more scenarios, or more domain-specific warnings.

5.2 Issues Arising from Diversities Across Regions and Cultures

The issues discussed in the previous section can arise even within a single society or culture. There is, however, another class of safety problems that appears when AI is expected to operate across different regions, cultures, or social contexts. In such cases, the same action, expression, or decision may be regarded as acceptable in one setting and inappropriate in another. Safety and ethical judgment therefore cannot always be treated as if they were uniform across all places and circumstances.

This problem is not limited to abstract values. It can also appear in ordinary practices and in the interpretation of familiar rules. A traffic marking, a signal, or a warning sign may have a

different practical meaning depending on the region in which it is used. Likewise, forms of greeting, acceptable physical distance, table manners, or the meaning of leaving food on a plate may vary across cultures. Even where people broadly agree that safety and consideration matter, the way these ideas are expressed in daily behavior can differ in ways that are significant for AI systems acting in the world.

For that reason, ensuring that an AI system is aligned with the expectations of one cultural or regional setting is not sufficient by itself. A system that behaves appropriately in one place may fail to do so elsewhere, and may even become intrusive or unsafe when moved into a different context. The challenge is not only to avoid technical errors, but also to avoid applying the wrong social assumptions to the situation at hand.

In the near term, one practical response is to adapt models or systems to the region or cultural setting in which they are used. Over the longer term, however, a more demanding goal emerges: AI should be able to detect when the surrounding context differs from the one it has learned, and either adjust its behavior accordingly or refrain from making a judgment when safe adaptation is not possible. In this sense, differences across cultures and regions are not a peripheral issue, but a central part of what it means for AI to act safely in the world.

5.3 Diverse Possible Safety Control Measures

The issues discussed so far cannot always be resolved within the AI model itself. Some of them depend on conditions outside the model, including the way the system is deployed, the way people use it, and the broader social environment in which it operates. For that reason, the question is not only how to improve AI technically, but also how responsibility for safety should be distributed across technical measures, organizational arrangements, and wider social responses.

Technical measures remain essential. They include methods for improving robustness against disturbed or degraded input, detecting manipulated or misleading content, and reducing obvious failure modes in model behavior. Such measures are often the first line of defense, and in some cases they may be sufficient. However, there are many situations in which technical measures alone cannot guarantee safe outcomes, even when they are well designed and carefully implemented.

A second layer of response lies in organizational arrangements. Safety may depend on who is allowed to rely on the system, how outputs are reviewed, how responsibilities are assigned, and what procedures are followed when the model is updated or when unusual behavior is observed. Problems such as misuse, overreliance, unclear authority, or failure to re-evaluate the system after change are not purely technical problems, even though they may interact with technical ones. They arise from the way the system is embedded in practice.

A third layer lies beyond any single organization. Some risks cannot be managed by developers or deployers alone because they involve other actors, public infrastructures, or the wider circulation of AI-generated material. Malicious use is one example: even if a developer improves the safety of one system, an attacker may simply use a different system. Similar difficulties arise when harmful outputs spread through public channels, when multiple organizations share responsibility without clear boundaries, or when social systems are not prepared for the scale and speed of AI-enabled activity. In such cases, technical measures must be connected with organizational rules and, in some cases, with broader social or regulatory arrangements.

The central challenge, therefore, is not merely to strengthen one kind of measure in isolation. It is to determine which kinds of problems should be handled mainly through technical design, which require organizational discipline, and which demand coordination at a wider social level. Without such a division of roles, safety work may become stronger in one place while remaining ineffective overall.

5.4 Preparing for Sudden Collaption of Technical Assumptions

The issues discussed in the previous sections are all framed in relation to AI as it exists today. They concern limitations that are already visible, differences that already affect judgment, and the ways in which responses must already be shared across technical, organizational, and social layers. There is, however, another kind of challenge. The assumptions on which current safety work is built may themselves become unstable as AI technology continues to change.

At present, much of the discussion around multimodal and agentic AI still depends on architectures derived from large language models. This makes it reasonable, for now, to study safety problems and countermeasures in ways that are closely tied to those architectures. In practice, this means that many current lines of research focus on phenomena that arise from the

design, training, and deployment patterns of today's systems. Such work remains necessary because the systems already being developed and deployed are built on these foundations.

At the same time, there is no guarantee that these foundations will remain dominant. AI has already changed rapidly within a short period, and it is possible that future architectures will alter the nature of the problems that safety research is trying to address. Measures that are effective under present assumptions may become less relevant, or even inadequate, if the mechanisms producing model behavior change significantly. A safety strategy that is too tightly bound to today's technical structure therefore carries the risk of becoming obsolete when that structure is replaced.

For this reason, safety research must proceed with two perspectives at once. One is to continue addressing the concrete problems posed by current AI systems, because those problems are real and immediate. The other is to prepare for discontinuous changes that may alter the technical basis of AI itself. This requires not only improving current countermeasures, but also examining which parts of present safety knowledge depend on temporary technical assumptions and which might remain valid across future shifts in architecture.

The discussion so far can be organized along four axes. The first concerns safety limits that become more serious as AI systems expand their range of action, including unstated assumptions, the limits of case-based learning, and the instability of safety knowledge when action extends across domains. The second concerns additional difficulty created by differences across regions, cultures, and social contexts. The third concerns the fact that meaningful responses cannot always be confined to technical measures, but must often be distributed across technical design, organizational practice, and broader social arrangements. The fourth concerns the possibility that the assumptions underlying current safety work may themselves change as AI technology evolves.

These four axes should not be treated as independent lists of concerns. In practice, several of them may be present at the same time, and safety work must often address their interaction rather than only one of them in isolation.

5.5 Research Agenda Catalog

The catalog below groups the research topics under the same four axes discussed in Sections 5.1–5.4.

5.5.1 Expanding Scope of Action

(1) Formalization and Articulation of Unstated Assumptions

Background: Many safety- and ethics-related assumptions in everyday life remain unwritten and therefore inaccessible to AI systems that rely on explicitly documented knowledge.

Core issue: The challenge is not the absence of knowledge, but the lack of representations that allow AI systems to isolate, interpret, and reuse such tacit assumptions.

Expected outcome: Methods for identifying, structuring, and representing implicit assumptions as reusable knowledge assets for AI design, evaluation, and operation.

Evaluability: Assessed by the coverage and consistency of extracted assumptions and the degree to which they improve safety-relevant judgment in practice.

(2) Machine-Readable Restructuring of Normative Documents

Background: Laws, standards, and guidelines are written in natural language and include implicit background interpretation, exceptions, and cross-references that AI cannot reliably process.

Core issue: The primary difficulty is not the absence of rules, but the absence of representations that enable stable and context-sensitive application by AI systems.

Expected outcome: Representation schemes for transforming normative documents into machine-readable forms that preserve applicability conditions, exceptions, and inter-rule relationships.

Evaluability: Evaluated by fidelity to the original normative meaning and behavioral comparison between systems with and without such structured representations.

(3) Analysis of AI's Ways of Understanding of Safety and Ethics

Background: How AI models internally represent safety and ethical concepts—and where such representations break down—remains insufficiently understood.

Core issue: Understanding the correspondence between internal model representations and safety-related concepts to identify failure points.

Expected outcome: Analytical methods that reveal internal structures relevant to safety/ethics and techniques to locate origins of unsafe behaviors.

Evaluability: Measured by explanatory coverage of failure patterns and the reduction of defects after targeted interventions.

(4) Addressing Insufficient Understanding of Real-World Phenomena

Background: Current AI relies heavily on training examples and struggles with phenomena not captured in data or with ambiguous real-world conditions.

Core issue: Whether AI can generalize and make safe decisions in situations not represented in its training data.

Expected outcome: Models that abstract physical and social regularities to support safe decision-making in unseen scenarios.

Evaluability: Evaluated through performance in unfamiliar scenarios and reduction in unsafe deviations.

(5) Generalization to Overcome Limits of Scenario Enumeration

Background: Real-world scenarios cannot be exhaustively enumerated, making case-based learning insufficient for safety.

Core issue: The need to move from data coverage toward structural and causal world understanding.

Expected outcome: Generalization mechanisms enabling safe behavior based on underlying principles rather than accumulated cases.

Evaluability: Assessed by behavioral consistency in novel contexts and reduction of unsafe extrapolation.

(6) Connecting Case-Based Learning with Principle-Based Understanding

Background: Case-derived behaviors and underlying principles are weakly integrated, limiting the robustness of safety behavior.

Core issue: Insufficient linkage between example-driven behavior and fundamental world understanding.

Expected outcome: Models capable of applying principle-based reasoning without excessive dependence on examples.

Evaluability: Measured by reduced case dependence and improved consistency of principle-driven decisions.

(7) Overcoming Limitations of Domain-Specific Safety Measures

Background: Safety designs are typically domain-specific, but AI increasingly crosses domain boundaries.

Core issue: Domain-restricted safety knowledge fails in boundary-crossing or mixed-context situations.

Expected outcome: Cross-domain safety principles and frameworks for dealing with unstable or hybrid environments.

Evaluability: Evaluated by reductions in safety failures in multi-domain scenarios.

(8) Detection and Suppression of Out-of-Scope Behavior

Background: AI may attempt actions beyond intended scope, and current detection/containment mechanisms are insufficient.

Core issue: Real-time detection of scope deviation and safe fallback control.

Expected outcome: Deviation early-warning models and integrated fallback safety mechanisms.

Evaluability: Measured by detection accuracy, false positives, and safe-fallback success rate.

(9) Self-Understanding and Judgment Deferral

Background: AI often fails to recognize its own uncertainty and produces unsafe judgments instead of deferring.

Core issue: Lack of mechanisms for uncertainty self-assessment and judgment avoidance.

Expected outcome: Meta-cognitive mechanisms enabling deferral or safe handoff when uncertainty is high.

Evaluability: Evaluated by appropriate deferral rates and reductions in erroneous judgments.

5.5.2 Regional and Cultural Differences

(10) Systematization of Regional and Cultural Differences

Background: Safety expectations and appropriate behavior vary across cultures and regions, but these differences lack systematic organization.

Core issue: Need for general frameworks rather than ad-hoc cultural examples.

Expected outcome: Structured models of cultural/regional differences enabling context-sensitive adaptation.

Evaluability: Measured by reduced misalignment and improved cross-cultural behavioral consistency.

(11) Detection of Contextual Misalignment and Adjustment of Safety Behavior

Background: AI may fail to notice when its assumptions conflict with local culture or situational context.

Core issue: Lack of mechanisms to detect contextual mismatches and shift toward safer alternatives.

Expected outcome: Context-awareness models enabling mismatch detection, judgment deferral, or safe reversion.

Evaluability: Evaluated by mismatch detection accuracy and the proportion of safe fallback behaviors.

(12) Judgment Principles for Handling Cultural and Contextual Differences

Background: Cultural differences cannot be managed through rule translation alone.

Core issue: Absence of abstract judgment principles that generalize safely across cultures.

Expected outcome: Universal safety-oriented judgment principles adaptable across cultural contexts.

Evaluability: Measured by decision stability across cultures and reduction of misunderstandings.

5.5.3 Diverse Possible Control Measures

(13) Integration of Technical, Organizational, and Societal Measures

Background: Risks from misuse, overreliance, unclear responsibilities, and unprepared societal systems cannot be addressed through technical measures alone.

Core issue: Lack of coordinated mechanisms for distributing safety responsibilities across different layers.

Expected outcome: Multi-layer response models combining technical safeguards, organizational processes, and social governance.

Evaluability: Evaluated by reductions in multi-factor incidents and improvements in system operation.

5.5.4 Sudden Collapse of Technical Assumptions

(14) Robustness Against Emergence of New Architectures

Background: Future AI architectures may diverge from current ones, making existing safety measures obsolete.

Core issue: Safety measures tied too tightly to current assumptions risk losing validity.

Expected outcome: Principle-based safety requirements applicable across architectural shifts.

Evaluability: Measured by cross-architecture applicability and sustained safety performance.

(15) Designing Research and Investment for Discontinuous Technological Shifts

Background: Short-term optimization can hinder preparedness for long-term, discontinuous shifts in AI technology.

Core issue: Over-fitting safety measures to current technology may cause future vulnerabilities.

Expected outcome: Research portfolios and investment strategies resilient across multiple future scenarios.

Evaluability: Evaluated by robustness and adaptability across divergent technological futures.

References

[1] Clifton A. Ericson II, *Hazard Analysis Techniques for System Safety, Second Edition*, Wiley, July 2015 ISBN: 978-1-118-94038-9

[2] Nancy G. Leveson, *Safeware: system safety and computers*, Association for Computing Machinery, April 1995 https://doi.org/10.1145/202709

[3] Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., & Thompson, N. (2024). *The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks from Artificial Intelligence*.
https://doi.org/10.48550/arXiv.2408.12622

[4] National Institute of Standards and Technology, NIST-AI 600-1, *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile,* July 2024
https://www.nist.gov/itl/ai-risk-management-framework

[5] Yoshua Bengio et. al., *International AI Safety Report 2025*, January 2025
https://internationalaisafetyreport.org/

[6] AI Security Institute, *The UK AI Security Institute's Research Agenda*, May 2025
https://www.aisi.gov.uk/research-agenda

[7] Japan AI Safety Institute, *Report on the Specific Influence on AI Safety*, October 2025
https://aisi.go.jp/output/output_information/251031/

Acknowledgements

This report was developed by Task Force 1 of the Committee for Machine Learning Quality Management, which consists of industry and academic experts organized by the National Institute of Advanced Industrial Science and Technology (AIST). The task force relied heavily on the discussions from the Summer Camp on the AI Safety 2025, which was also organized by AIST. The report was developed as part of a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

The members of Task Force 1 of the Committee for Machine Learning Quality Management (in alphabetical order of surname) are as follows:

Takashi EGAWA	AIST
Chinami HAMATANI	Ad-sol Nisshin Corporation
Masahiro HAYASHITANI	NEC Corporation
Yoshinao ISOBE	AIST
Yusuke KAWAMOTO	AIST
Takashi KITAMURA	AIST
Koichi KONISHI	AIST
Kazumasa MIYAKE	Sumitomo Electric Industries, Ltd.
Takeshi MIYAKE	Cyber Creative Institute Company Limited
Shin NAKAJIMA	AIST / National Institute of Informatics (NII)
Takaaki NAMBA	AIST
Yutaka OIWA	AIST
Tamao OKAMOTO	SHIN-JIGEN Inc.
Yoshiki SEO	AIST
Akira TANAKA	AIST

The participants in the Summer Camp on the AI Safety 2025 are as follows (in alphabetical order of surname, with affiliations shown as of August 1, 2025):

Chinami HAMATANI	Ad-sol Nisshin Corporation
Masahiro HAYASHITANI	NEC Corporation
Hasumi ICHIKAWA	Ernst & Young ShinNihon LLC
Taruho INADA	LY Corporation
Shinji ISHII	Cyber Creative Institute Company Limited
Kazunobu ISHIKAWA	Panasonic Holdings Corporation

Yoshinao ISOBE	AIST
Yusuke KAWAMOTO	AIST
Hiromu KITAMURA	Japan AI Safety Institute (J-AISI) / Information-technology Promotion Agency, Japan (IPA)
Takashi KITAMURA	AIST
Koichi KONISHI	AIST
Takeshi MIYAKE	Cyber Creative Institute Company Limited
Kazumasa MIYAKE	Sumitomo Electric Industries, Ltd.
Junichi MIYAMOTO	New Energy and Industrial Technology Development Organization (NEDO)
Shin NAKAJIMA	AIST / National Institute of Informatics (NII)
Takaaki NAMBA	Panasonic Holdings Corporation
Leo OHTSUKA	Japan Quality Assurance Organization (JQA)
Yutaka OIWA	AIST
Tamao OKAMOTO	SHIN-JIGEN Inc.
Yosuke ONO	Japan Quality Assurance Organization (JQA)
Yoshiki SEO	AIST
Yoshifumi SHIBATA	New Energy and Industrial Technology Development Organization (NEDO)
Tatsuo SUGA	Ernst & Young ShinNihon LLC
Asei SUGIYAMA	Citadel AI Inc.
Akira TANAKA	AIST
Masahiro TANAKA	National Agriculture and Food Research Organization (NARO)
Hiroki TSUKIDA	Forcesteed Robotics
Teresa TSUKIJI	SherLOCK, Inc.
Tomoyuki UENO	Corpy & Co., Inc.
Yoshinori YANAGAWA	FUJIFILM Corporation
Mahoko YASUE	New Energy and Industrial Technology Development Organization (NEDO)