

By



Hiroki Habuka

Representative Director & CEO of Smart Governance Ltd. Research Professor, Graduate School of Law, Kyoto University Visiting Professor, University of Tokyo Graduate School of Law and Politics. Attorney-at-Law (Japan & New York State)

Hiroki specializes in law, corporate governance, and social governance in the context of the AI- and data-driven society. He holds LL.B. and J.D. degrees from the University of Tokyo, and an LL.M. from Stanford Law School as a Fulbright Scholar. His professional experience includes practice at Mori Hamada & Matsumoto, as well as service at the Financial Services Agency of Japan and the Ministry of Economy, Trade and Industry (METI). During his tenure at METI, he was the principal drafter of the Ministry's report on Agile Governance. In 2020, he was named by the World Economic Forum as one of the "Top 50 Most Influential People Transforming the Public Sector." His major publication is Introduction to AI Governance: From Risk Management to Social Design (Hayakawa Publishing). He concurrently serves as a Senior Science and Technology Policy Fellow at the Secretariat of the AI Strategy Headquarters (Cabinet Office, Government of Japan), the Representative Director of the Japan Association for AI Governance, Visiting Professor at the University of Tokyo Graduate School of Law and Politics, Non-Resident Fellow at the Center for Strategic and International Studies (CSIS), member of the World Economic Forum's Global Future Council on GovTech and Digital Public Infrastructure.

Explore more from
The Kyndryl Institute
kyndryl.com/institute

Agile governance in the age of frontier AI

According to research, 52% of business leaders worldwide say that risks are now growing faster than they can address them. While 75% say their internal governance is effective today, only 28% believe their organizations are adequately prepared for future risks.

For CEOs, CIOs, risk leaders, and policymakers, the challenge is how to govern technologies whose future capabilities and risks cannot be fully predicted. The leadership response is agile governance: governance systems that continuously monitor, learn, adapt and evolve alongside technological change.

The world today is filled with uncertainty. In particular, 2026 was the year frontier AI — cutting-edge AI that is not merely highly capable but may give rise to capabilities and risks that have never existed before — moved from a concept discussed in policy and research circles to a tangible reality. In April, Anthropic introduced its advanced AI model Claude Mythos as "a model considered too risky for public release." On the same day, it launched Project Glasswing, an initiative under which AI model developers and critical infrastructure operators would test the model in advance. This represents a new form of governance: deployment proceeds in

tandem with risk assessment. Additionally, in July, an internal evaluation conducted by OpenAI revealed a case in which an AI agent bypassed its intended containment measures and, while coordinating with other agents, gained access to systems operated by Hugging Face, an external platform. We have therefore entered an era in which AI itself helps develop AI, making it impossible to anticipate every risk and manage it through fixed rules alone. What becomes essential is agile governance: an approach that assumes failures and unforeseen events will occur and continually updates rules and operations through ongoing monitoring and learning. The question is no longer how to create perfect rules. The question is how to build organizations capable of continuously learning and adapting.

Why static governance falls short

The EU AI Act was designed as the world's first law to regulate AI across the board. When the original proposal was published in 2021, it was designed as a “future-proof” framework capable of accommodating technological change. At the time, however, the legislation primarily envisioned purpose-specific AI, such as image-recognition systems and recommendation algorithms. The arrival of ChatGPT in November 2022 made clear that the original framework could not adequately address the new technology. A new chapter on General-Purpose AI was ultimately added to cover generative AI. This illustrates how difficult it is to govern technological innovation through rules designed in advance.

The EU AI Act sets out fundamental principles and obligations, while leaving many of the specific implementation methods to European harmonized standards that are still to be developed. Yet the risks and required countermeasures differ greatly among chatbots, industrial AI, medical AI, and financial AI. Consequently, developing standards and guidance that both establish common requirements and address the differing characteristics of individual use cases is far from straightforward. Reflecting these implementation challenges, the entry into application of major obligations for high-risk AI systems

under the EU AI Act has been delayed by more than one year.

Japan, by contrast, made its position clear: to encourage AI, it would adopt a regulatory approach designed to be as flexible as possible. The theoretical foundation for this approach is the Governance Innovation Report series published by Japan's Ministry of Economy, Trade and Industry (METI) since 2020. I served as a principal author of the first three editions of these reports. When we were preparing the first report, however, I believed I was writing about a future more than a decade away. Since the emergence of ChatGPT, change has accelerated dramatically. Within enterprises, software development, customer service, content creation, and many other activities have been transformed in a short period. As a result, responding flexibly to such change has become an urgent priority across society. Against this backdrop, the AI Promotion Act, enacted in May 2025, is — as its name suggests — a framework intended not to restrict AI use but to promote it. Reflecting the principles of agile governance, it calls for the continuous monitoring of AI adoption, risks, and mitigation measures, with policies reviewed promptly in response. In doing so, it clearly commits the government itself to ongoing situational assessment and policy updates.

The contrast between the EU and Japan illustrates a broader lesson. The future of AI governance will not be determined by who writes the most

We have therefore entered an era in which AI itself helps develop AI, making it impossible to anticipate every risk and manage it through fixed rules alone. What becomes essential is agile governance: an approach that assumes failures and unforeseen events will occur and continually updates rules and operations through ongoing monitoring and learning.

comprehensive rules. It will be determined by who builds the greatest capacity to learn and adapt as technology evolves.

themselves, including employment systems, democracy, and education, should adapt to an AI-driven future.

What are the risks posed by AI?

So what does AI risk ultimately mean? As AI capabilities expand across domains as diverse as those in which humans operate, the risks AI may create are also extremely varied (just as it is difficult to give an immediate answer when asked, “What risks do humans pose?”). Broadly speaking, however, these risks can be grouped into three categories.

1. Malfunction

These are risks arising from improper AI behavior, including hallucinations, outputs containing inappropriate bias, and AI agents acting beyond their intended boundaries. Developers must continually work to improve safety. At the same time, users must proceed on the assumption that AI can make mistakes and establish organizational mechanisms for oversight and verification.

2. Misuse

These are risks caused by the malicious use of AI, including cyberattacks, deepfake fraud, the spread of disinformation, and applications in weapons development. The key responses are to strengthen defenses at the organizational and societal levels and for frontier AI developers to build guardrails that prevent harmful outputs. Because restricting the development of cyber capabilities can also limit defenders’ capabilities, rigorous access-rights management is equally important.

3. Systemic risk

These are structural risks that emerge as AI becomes deeply embedded throughout society. Examples include effects on employment, the erosion of human agency, the emergence of a surveillance society, and impacts on democracy. These risks cannot be addressed simply by improving the safety of individual AI systems. Governments, businesses, and civil society must work together to consider how social institutions

The importance of distinguishing among malfunction, misuse and systemic risk extends beyond classification. Each category requires a different response, which is precisely why leaders need agile governance. Static frameworks seek one answer to every risk. Agile governance creates mechanisms for continuously identifying which response is needed, when, and by whom. Even the phrase “AI cyber risk” evokes different concerns for different people. Some envision AI systems malfunctioning or being hijacked after a cyberattack. Others worry about AI agents going out of control and penetrating external systems. Still others focus on attackers using AI to conduct sophisticated cyberattacks. If people are discussing different kinds of “risk,” the conversation will inevitably fail to align. Organizations must therefore begin by defining clearly what they consider to be a risk.

In recent years, a growing number of frontier AI developers and operators have begun presenting evidence on the risks associated with their AI systems, the measures implemented to mitigate those risks, and the extent to which those measures have been effective. These developments are not limited to voluntary efforts by companies. Regulators are also beginning to incorporate similar approaches into policy frameworks. In the European Union, providers of

The future of AI governance will not be determined by who writes the most comprehensive rules. It will be determined by who builds the greatest capacity to learn and adapt as technology evolves.

general-purpose AI models are required to maintain technical documentation, while developers of particularly high-risk models must conduct evaluations and report serious incidents. In California, legislation has been enacted requiring frontier AI developers to disclose their safety and risk-management frameworks. In Japan, discussions have likewise begun on whether the government's information-gathering powers should be strengthened. Developers and operators possess far more detailed knowledge than regulators about the technology and the actual nature of the risks. For that reason, mechanisms are needed through which operators proactively assess risks and share their findings with society and regulators, enabling safety to improve continuously. This is also the essence of agile governance. Businesses, users, governments, researchers, and other stakeholders bring together their respective expertise and update rules and operations as circumstances change. Only by sustaining this dialogue among stakeholders can society advance responsible innovation as a whole.

Questioning existing rules and systems

One of the core concepts of agile governance is the “double feedback cycle.” The first form of feedback assesses whether frontline operations are being conducted appropriately and whether misconduct or other problems have occurred. The second, equally important form of feedback examines whether the organization's rules, mission, and staffing structure remain suited to the current environment. Rules that were once optimal may cease to function as society and technology change. Organizations must therefore do more than monitor whether employees follow the rules; they must continually question whether the rules themselves remain effective. Continuous feedback at both levels — operational improvement on the front line and management-level review of institutions and organizational structures — constitutes the risk management needed to navigate an uncertain society.

For AI agents, for example, the two cycles can be understood as operating in tandem: one continuously evaluates and improves the agent's behavior

and operation, while the other reviews the governance framework itself — including which tools the AI may access, how much authority it should be granted, and under what circumstances a human should intervene.

Human well-being as the ultimate goal

The use of AI will remain unavoidable. Because using AI inevitably creates risks, adoption and risk management must be considered together. Where should companies begin? I would point to two priorities: taking inventory of AI use and securing the necessary talent. The first step is an inventory — understanding who within the organization is using AI and where. “Shadow AI,” in which employees use AI for work through personal accounts, has become a concern, but regulation alone will not be effective. Companies must also consider how to provide an environment in which employees can use AI easily and appropriately.

Securing talent does not necessarily mean hiring new people from outside. AI governance encompasses legal affairs, quality management, privacy, intellectual property, branding, communications, and many other areas. Each function sees different issues, anticipates different risks, and has a different scope of responsibility. Creating a forum where people with diverse expertise can bring their knowledge together, treat AI as an issue that concerns them directly, and discuss it collectively will be critical and will support enterprises in being ready for the future. Ideally, this process will help the organization share a common understanding of its ultimate goals: “What do we want to achieve with AI?” and “What do we want to protect?”

Agile governance as a competitive leadership capability

AI adoption and risk management can no longer be treated as separate conversations. Organizations that move fastest without effective governance will create unacceptable risk. Organizations that focus only on control may fail to capture the benefits of

innovation. The challenge for leaders is therefore not choosing between innovation and governance but combining them.

Agile governance offers a way forward. By continuously monitoring outcomes, questioning assumptions, updating rules and engaging diverse stakeholders, organizations can remain resilient amid uncertainty while still pursuing innovation. Ultimately, the purpose of governance is not to preserve existing systems. It is to support human agency and well-being in a changing world. As AI transforms society, leaders must ensure their organizations remain capable of adapting alongside it. —

Agile governance offers a way forward. By continuously monitoring outcomes, questioning assumptions, updating rules and engaging diverse stakeholders, organizations can remain resilient amid uncertainty while still pursuing innovation.

